Comparaison de groupes de variables définies sur le même ensemble d'individus

Brigitte Escofier, Jérome Pagès

IRIA

CENTRE DE RENNES
IRISA

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P 105
78153 Le Chesnay Cedex
France
Tél.: 954 90 20

Rapports de Recherche

N° 149

COMPARAISON DE GROUPES DE VARIABLES DÉFINIES SUR LE MÊME ENSEMBLE D'INDIVIDUS

Brigitte ESCOFIER
Jérôme PAGES

Juillet 1982

INRIA

CENTRE DE RENNES
IRISA

Rapports de Recherche

N° 149

COMPARAISON DE GROUPES DE VARIABLES DÉFINIES SUR LE MÊME ENSEMBLE D'INDIVIDUS

Brigitte ESCOFIER
Jérôme PAGES

Juillet 1982

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél.: 954 90 20

IRISA

PUBLICATIONS INTERNES

COMPARAISON DE GROUPES DE VARIABLES DEFINIES SUR LE MEME ENSEMBLE D'INDIVIDUS

Brigitte ESCOFIER
Jérôme PAGES

Publication Interne n°166 - Mai 1982

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE
ET SYSTEMES ALEATOIRES

Campus Universitaire de Beaulieu
Avenue du Général Leclerc
35042 - RENNES CÉDEX
FRANCE
TéL. : (99) 36.20.00
Télex : UNIRISA 95 0473 F

COMPARAISON DE GROUPES DE VARIABLES DEFINIES SUR LE MEME ENSEMBLE D'INDIVIDUS

Brigitte ESCOFIER*

Jérôme PAGES **

Publication n° 166 - 121 pages - Mai 1982

Résumé : Lorsque l'on veut étudier un tableau "individus x variables" dans lesquel les variables sont structurées en groupes, plusieurs problèmes particuliers se posent. Comment équilibrer l'influence des différents groupes dans une analyse factorielle ? Ces groupes induisent-ils les mêmes structures sur l'ensemble des individus ? Existe-t-il un moyen graphique pour comparer ces structures ? Peut-on mettre en évidence les liaisons entre les différents groupes de variables, en extraire : des facteurs communs ? Quels sont les groupes de variables qui se ressemblent le plus ?
Ces questions ont déjà suscité des travaux qui ont conduit à plusieurs méthodes, l'analyse multicanonique (CARROLL), statis (ESCOUFIER).
Nous proposons une nouvelle méthode qui répond simultanément et de manière optimale (en un certain sens) à toutes ces questions. Les résultats sont basés sur une analyse factorielle du tableau dans laquelle chaque groupe de variable est pondéré. Ils permettent une représentation géométrique : des individus caractérisés par l'ensemble de toutes les variables ; des individus caractérisés par chacun des groupes de variables ; des variables et des composantes principales de chaque groupe ; des groupes de variables.

Abstract : The factor analysis of the "individuals x variables" table in which the variables are structured in groups, presents many particular problems such as :
Do the variable groups induce the same structures on the set of individuals ? Do there exist graphicaal methods to compare these structures ?
These questions have given rise to many works deading to different methods, for instance multicanonical analysis (CARROLL), statis (ESCOUFIER). We propose a new method which has an interesting property of deading to the computations geometrically interpretable in the three spaces that are classical used in this type of problem :
observation space, variable space and the space of variable groups.

* IRISA
** ENSAR, route de St. Brieuc Rennes

Secrétariat de rédaction : Melle F. MOINET - Lab. d'Informatique
Campus de Beaulieu
35042 RENNES CEDEX

SOMMAIRE

INTRODUCTION

Les techniques d'analyse des données permettant d'étudier un seul tableau de données sont maintenant classiques.

Mais depuis plusieurs années, les statisticiens sont de plus en plus souvent consultés, aussi bien par des chercheurs de sciences de la nature que des sciences humaines à propos de problèmes impliquant l'étude simultanée de plusieurs tableaux. Il s'agit souvent de suites de tableaux indicés par le temps, ou de tableaux provenant d'un unique tableau de dimension trois, mais on rencontre aussi fréquemment des ensembles de tableaux définis par différents groupes de variables mesurées sur les mêmes individus ou différents groupes d'individus caractérisés par la même variable etc ...

Ces problèmes s'inscrivent en droite ligne de l'évolution des quinze dernières années, permise par le développement de l'analyse des données, lui-même permis par celui de l'informatique : au raisonnement "toutes choses égales par ailleurs", au découpage cartésien d'un problème complexe en multiples problèmes plus simples, on substitue souvent une approche globale visant à prendre en compte simultanément tous les aspects d'un même phénomène.

Dans cet article nous nous restreindrons à un cas particulier d'ensemble de tableaux. Tout d'abord, chaque tableau met en correspondance un ensemble d'individus ou un ensemble de variables (en rassemblant les valeurs prises par chacune des variables à propos de chaque individu). Ensuite, l'ensemble des individus est le même pour tous les tableaux.

Les raisons qui nous ont conduits à cette restriction sont les suivantes : Les questions que l'on peut se poser devant un ensemble de tableaux diffèrent suivant la nature de ces tableaux. Dans le cas étudié ici, la notion de liaison entre groupes de variables est une notion fondamentale que nous étudions longuement. Cette notion ne se traduit pas directement dans d'autres cas.

Ensuite ce type de tableau est très fréquemment rencontré. Il recouvre lui-même plusieurs situations selon que les variables sont numériques ou qualitatives. Précisons qu'une variable qualitative définit une partition de l'ensemble des individus et qu'une telle variable est représentée classiquement en analyse factorielle par l'ensemble des variables indicatrices de toutes les classes de cette partition.

Remarquons au passage les cas très particuliers où chaque groupe de variables ne comprend qu'une variable numérique ou qualitative. Ils constituent des cas limites auxquels il sera intéressant de se référer. Dans le contexte d'analyse factorielle dans lequel nous nous placerons principalement, on sait traiter simultanément des groupes d'une seule variable, par l'analyse en composantes principales s'il s'agit de variables numériques et par l'analyse des correspondances multiples s'il s'agit de variables qualitatives. Mais bien entendu, ce sont les cas non triviaux où chaque groupe de variables comporte plusieurs variables numériques ou plusieurs variables qualitatives qui nous intéresserons véritablement.

Toutefois, les résultats exposés ici et les méthodes d'analyse proposées peuvent s'adapter à l'étude simultanée d'autres ensembles de tableaux. Nous l'exposerons ultérieurement.

L'objet de ce travail est d'abord de détailler les questions que l'on peut se poser devant un ensemble de tableaux regroupant des variables définies sur les mêmes individus. Nous le ferons en référence à des méthodes connues dont les objectifs peuvent sembler très divers. (analyse canonique généralisée, méthode STATIS, utilisation des opérateurs dit d'Escoufier, méthode procustéenne, modèle INDSCAL).

Nous verrons que toutes ces questions peuvent être formulées en termes de trois types d'objets : les individus, les variables, les groupes de variables.

La deuxième étape est une présentation des divers espaces dans lesquels on représentera ces trois types d'objets. Ces espaces constituent des cadres de référence utiles pour formaliser mathématiquement les questions préalablement posées et les solutions proposées. Ces cadres ne sont pas nouveaux. La plupart des représentations des objets sont tout à fait classiques, mais certaines sont originales. Nous décrivons leurs propriétés et les relations qui les lient d'une façon plus complète qu'il n'est fait habituellement.

Dans la troisième étape, nous proposons une méthode nouvelle de traitement simultané de groupes de variables . Cette méthode permet d'obtenir des représentations factorielles des variables, des groupes de variables, de l'ensemble des individus vu à travers chaque groupe de variables et vu à travers l'ensemble de toutes les variables, toutes ces représentations étant liées par des relations précises. Cette méthode s'interprète dans les différents cadres de référence, ce qui fait qu'elle a plusieurs facettes et qu'elle permet de répondre de façon cohérente aux diverses questions posées dans la première étape.

Le simple fait que cette interprétation soit possible dans tous les cadres nous parait un argument important en faveur de la méthode.

Les interprétations dans les différents cadres rendront aisée la comparaison de cette méthode avec d'autres antérieurement proposées. Tel est l'objet de la quatrième étape. Dans le cas où les groupes sont réduits à une seule variable qualitative ou quantitative, on obtient les résultats -un peu enrichis- de l'analyse en composantes principales ou de l'analyse des correspondances multiples.

1. LES OBJECTIFS SOUS-JACENTS A L'ETUDE SIMULTANEE DE PLUSIEURS GROUPES DE VARIABLES

1.1. Notations

A chaque groupe de variables correspond un tableau. Tous les groupes de variables étant définis sur le même ensemble d'individus, tous les tableaux peuvent être juxtaposés et former ainsi un tableau unique croisant l'ensemble des individus et l'ensemble des groupes de variables. L'ensemble initial de plusieurs tableaux apparaît alors comme un unique tableau structuré en sous tableaux.

Exploitant ce point de vue, nous appelons :

X le tableau complet ;

I l'ensemble des individus ;

K l'ensemble des variables (tous groupes confondus).

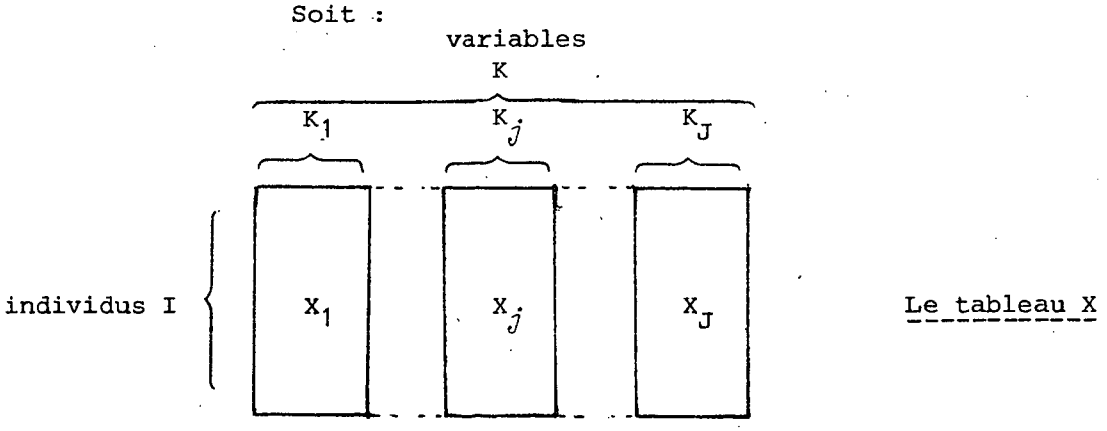
Le tableau X est structuré en sous-tableaux correspondant aux groupes de variables. On note :

J le nombre de sous-tableaux ;

K_j l'ensemble des variables du groupe j $\quad (K = \bigcup_j K_j)$

X_j le tableau associé au groupe j.

Soit :



Le tableau X

Selon un usage fréquent, les signes K, I ou K_j représentent à la fois un ensemble et son cardinal.

Une variable du groupe K_j est notée v_k, $k \in K_j$. Chaque variable v_k est affectée d'un poids, noté m_k. Aux individus, on affecte aussi des poids p_i, la somme des p_i est égale à 1.

1.2. Variables et structure sur les individus

Nous rappellerons tout d'abord la dualité qui existe dans l'étude d'un tableau de données tel que ceux dont nous disposons ici. L'analyse des individus (vus au travers des variables) et l'analyse des variables (mesurées sur les individus)sont deux approches du même phénomène : le tableau.

Or les données auxquelles nous nous intéressons se présentent sous la forme de groupes de variables (ou de tableaux) décrivant les mêmes individus. Il s'ensuit que pour étudier et comparer ces groupes de variables, nous pourrons choisir deux points de vue :

. soit considérer les variables ou les groupes de variables eux-mêmes,
. soit considérer les structures induites sur le même ensemble d'individus.

Nous donnons au mot"structure"son sens habituel en analyse factorielle. Ainsi chaque groupe de variables définit sur l'ensemble des individus une distance qui mesure leur ressemblance. La structure de l'ensemble des individus est alors l'ensemble de ces distances interindividuelles. Les individus étant classiquement représentés par un nuage de points dans un espace euclidien (voir § 2.2), la structure de l'ensemble est souvent appelée "forme du nuage".

1.3. Comparaison globale des groupes de variables

Un des objectifs poursuivis dans l'étude simultanée de plusieurs groupes de variables peut être une comparaison globale de ces groupes. Dans le contexte un peu plus général de la comparaison de tableaux, Y. ESCOUFIER et L'HERMIER DES PLANTES, intitulent ce problème "étude de l'interstructure"[cf9 et16]

Les objets que l'on cherche à comparer sont les tableaux ; les questions posées sont :

- les tableaux sont-ils très différents ou non ?
- Quels sont ceux qui se ressemblent, ceux qui sont différents..?
- Certains tableaux peuvent-ils être considérés comme "intermédiaires"entre deux ou plusieurs autres ?
- Peut-on présenter une typologie des tableaux ?

Il est clair que toutes ces questions reposent sur une mesure globale de liaison - ou de distance - entre les groupes de variables (ou les tableaux).

Nous introduirons et nous étudierons aux § 2-5-2 et 2-5-3 cette mesure de liaison. Notons seulement ici qu'il y a trois points de vue pour introduire une mesure de distance ou de liaison entre groupes de variables :

- généralisation des mesures de liaisons habituelles corrélation linéaire, multiple et canonique.
- mesure de ressemblance entre les structures induites sur le même ensemble.
- mesure de ressemblance entre les formes générales des nuages de variables.

Pour visualiser les distances entre tableaux, dans le contexte d'analyse factorielle qui est le nôtre, il est naturel d'en chercher une représentation approchée dans des espaces de petite dimension figurés par des plans. C'est ce que nous proposent Y. ESCOUFIER et L'HERMIER DES PLANTES dans leur étude de l'interstructure. Mais, dans cette étude, seuls les groupes de variables apparaissent, ce qui limite les possibilités d'interprétation. Des éléments explicatifs seraient très utiles pour utiliser ce type de graphique. Ici les éléments explicatifs peuvent être les variables qui composent les groupes ou bien les individu ou bien encore les composantes principales des groupes de variables. L'introduction de ces éléments permettrait de préciser en quoi les tableaux se ressemblent et en quoi ils diffèrent.

Pour comparer globalement les groupes de variables nous chercherons à définir une mesure de liaison entre ces groupes. Puis, nous chercherons une représentation approchée des distances entre les groupes étudiés complétée par des représentations des individus et des variables intervenant comme éléments explicatifs

1.4. Comparaison des nuages d'individus

Dans ce paragraphe, on s'intéresse aux différents nuages d'individus définis par chacun des groupes de variables et on souhaite réaliser une comparaison analytique de ces divers nuages. Le terme "analytique" est utilisé en opposition au terme global du paragraphe précédent car le but poursuivi est maintenant d'analyser et non plus de mesurer les différences et les ressemblances entre les nuages.

Cette analyse elle-même peut avoir plusieurs points de vue. On peut étudier l'évolution de chaque individu à travers les groupes de variables, c'est à dire l'évolution de ses distances avec les autres individus. C'est dans cette optique qu'est introduite ce que Y. Escoufier et L'HERMIER DES PLANTES appellent l'étude des "intra-structures". Pour cette comparaison des positions d'individus, la solution proposée est une représentation simultanée des nuages d'individus sur des plans ou dans un espace de petite dimension.

Etudions les problèmes posés par une telle représentation et les propriétés intéressantes qu'il serait souhaitable qu'elle possède.

1.4.1. Représentation simultanée des nuages d'individus

Pour cette représentation simultanée, deux problèmes se posent : celui de la réduction de chaque nuage sur un espace de petite dimension et celui de l'orientation commune des nuages. En effet, une ressemblance intrinsèque entre deux nuages peut être masquée par des symétries ou des rotations si on se contente de superposer les repères de nuages qui, a priori, sont situés dans des espaces différents (voir § 2.2). Trois approches sont possibles pour résoudre ces problèmes :

(1) réduction des dimensions de nuages, puis orientation des nouveaux repères,

(2) orientation commune des repères de chaque nuage, puis réduction des dimensions,

(3) orientation et réduction simultanées.

Pour comparer ces trois approches , cherchons quelles sont les qualités souhaitables d'une telle représentation, sachant que le but de cette représentation est la comparaison des positions des individus dans les différents nuages. Deux propriétés doivent être vérifiées pour que le résultat soit intéressant :

$[P_1]$ Chaque nuage doit être "bien représenté". Nous traduisons mathématiquement cette propriété en termes familiers en analyse factorielle : tout d'abord la représentation d'un nuage doit être une projection de ce nuage, et de plus cette projection doit avoir une inertie suffisamment importante par rapport à l'inertie totale du nuage.

$[P_2]$ Les représentations des différents nuages doivent se ressembler pour qu'une comparaison soit possible : les points représentant le même individu doivent être proches les uns des autres.

Il est clair qu'il faudra trouver un compromis entre ces deux propriétés :

Si on ne tient compte que de la première, les meilleures représentations de chaque nuage sur un plan sont les projections sur le plan engendré par ses premiers axes d'inertie. Ces projections ne se ressemblent pas forcément, une simple permutation dans l'ordre des axes 2 et 3 rendrait la comparaison impossible.

Si l'on ne tient compte que de la deuxième, on risque d'obtenir des projections des nuages dans des directions d'inertie très faible et donc de mauvaises représentations de ces nuages.

Ces réflexions nous amènent à conclure que la première approche du problème qui consistait à réduire les dimensions de chaque nuage, puis à orienter les repères privilégie la propriété (p_1) aux dépens de la ressemblance entre les nuages.

La seconde approche, qui consistait à orienter d'abord les repères, puis à réduire les dimensions du nuage, privilégie au contraire la ressemblance (p_2) entre nuages aux dépends de la qualité de leur représentation (p_1).

La première approche évoque une analyse en composantes principales de chaque groupe de variables et la superposition des résultats. La seconde approche évoque les techniques d'analyse procustéenne [cf 10 et 19]. A la troisième approche, qui risque de mieux répondre au problème que nous avons posé puisqu'elle cherche dès le départ un compromis entre les deux propriétés, la méthode STATIS de L'HERMIER DES PLANTES [cf 16] répond en partie, bien qu'elle ne soit pas du tout posée dans ces termes et surtout que la représentation des différents nuages d'individusn'en soit pas une projection. Cette dernière propriété nous parait très importante car elle assure une certaine qualité de représentation. Elle permet de plus de mesurer cette qualité de représentation suivant les critères habituels, ce qui est un élément important de l'interprétation des résultats.

1.4.1-a. Nuages moyens

Notons que l'on facilitera beaucoup la comparaison des nuages en construisant un nuage moyen ou "compromis" qui a la propriété de ressembler le plus possible à l'ensemble des nuages. En effet, on substitue alors aux comparaisons deux à deux (en nombre J x J), J comparaisons à une moyenne. Il est donc souhaitable que dans la représentation simultanée apparaisse ce nuage moyen.

Il est toujours possible de construire dans la représentation simultanée un nuage moyen au centre de gravité des différents nuages. Mais il serait beaucoup plus intéressant de construire a priori un nuage compromis entre les différents nuages et de le représenter ensuite avec les autres nuages par une projection orthogonale.

Dans STATIS la représentation simultanée contient, et est même fondéesur la représentation d'un nuage compromis. Mais, ce nuage compromis n'est pas situé au centre de gravité des différents nuages, ce qui aurait facilité comparaison. Nous chercherons une solution possédant cette propriété.

1.4.1-b. le repérage des individus

Le fait d'analyser des structures définies sur un ensemble d'individus suppose que ces individus ne sont pas trop nombreux et que leur identité est chargée de sens pour l'interpréteur. Or, les cas sont fréquents où ces deux conditions ne sont pas satisfaites. Dans cette situation, il convient de remplacer les individus par des classes d'individus définies à partir de variables qualitatives que l'on suppose être en rapport avec le phénomène que l'on étudie. C'est ainsi qu'à partir de données d'enquêtes dans lesquelles les individus sont initialement des enquêtés, on s'intéressera, par exemple, aux structures induites sur les tranches d'âge, les catégories socio-professionnelles...

1.4.2. Comparaison des formes des nuages d'individus

La comparaison des nuages peut se poser sous un aspect un peu moins individuel pour tenter de dégager en quoi les nuages se ressemblent et en quoi ils diffèrent. Les questions sont alors par exemple :

- Existe-t-il des projections des nuages sur des espaces de petite dimension qui se ressemblent ? En voyant ces projections comme des décompositions des distances interindividuelles, la question devient : Y-a-t-il des éléments de décomposition analogues ? Des facteurs de dispersion analogues ?

- Existe-t-il au contraire des directions de dispersion spécifiques d'un ou plusieurs nuages ?

- Si oui, lesquelles ?

Ces questions ne sont pas classiques, mais elles amènent à poser le même problème que ci-dessus :
la recherche de projections des différents nuages réalisant un compromis entre les propriétés (p_1) et (p_2). La propriété (p_2) (de ressemblance entre les projections) servira à dégager les structures communes, la propriété (p_1) (de qualité de représentation) à dégager des structures significatives. On proposera donc les mêmes solutions. Généralement les projections des différents nuages ne coincident pas. Pour mettre en évidence la structure commune, il est commode d'utiliser une représentation moyenne.
Le nuage moyen introduit au paragraphe précédent (1.4.1-a) sera bien adapté.

Mais le point de vue choisi ici amène à vouloir, si possible, compléter les projections simultanées des nuages par des éléments explicatifs. Citons-en quelques uns sans préciser leur forme exacte :

. un indice mesurant la ressemblance globale entre les représentations de tous les nuages, (ou la ressemblance de tous les nuages avec le nuage moyen). Ceci pour chaque axe de projection et éventuellement chaque plan.Cet indice permettra de juger si la représentation du nuage moyen peut être considérée comme une structure commune à tous les nuages. Si cet indice est trop faible, les groupes de variables sont peu liés, ou bien les ressemblances ont été extraites dans les axes précédents. Il sera alors vraisemblablement peu utile de comparer point par point ces projections.

. des indices mesurant, pour chaque groupe de variables la ressemblance entre la projection du nuage défini par ce groupe et la projection du nuage moyen. Ces indices permettront de préciser l'indice précédent en permettant de repérer les groupes possédant la structure commune et ceux qui ne la possèdent pas.

. des indices mesurant la qualité de représentation de chaque nuage. Ils permettront de mesurer l'importance des structures dégagées pour chaque nuage.

. des représentations des variables de tous les groupes pour mettre en évidence celles qui sont les plus liées aux structures dégagées, et donc responsables des distances interindividuelles visibles sur les graphiques.

Pour résumer, il faudrait une représentation simultanée des J nuages associés aux groupes de variables sur des espaces de petite dimension qui soient de bonnes représentations de ces nuages (p_1), qui se ressemblent (p_2), qui contiennent la représentation d'un nuage moyen et qui soit complétée par les indices ci-dessus.

1.5 Comparaison des nuages de variables

Les nuages représentant les variables sont constitués d'éléments différents mais sont situés dans le même espace puisque l'ensemble des individus est commun. La comparaison de ces nuages est un point de vue, qui, à notre connaissance n'a pas encore été étudié .

Les éléments caractéristiques des formes de ces nuages sont les directions de leurs axes d'inertie et les moments d'inertie associés. Ils sont plus intéressants que ces axes d 'inertie sont les composantes principales de chaque groupe et donc des éléments caractéristiques des tableaux. Pour comparer ces axes, on peut calculer les angles qu'ils font entre eux ou plutôt le cosinus de ces angles (i.e. la corrélation entre toutes les composantes principales). Mais le nombre de composantes rend longue cette étude. Pour faciliter cette comparaison, nous proposerons de projeter les vecteurs directeurs unitaires de ces axes sur des espaces de petite dimension,ce qui permettra de repérer facilement les composantes proches entre elles. Pour avoir une bonne représentation des angles, entre ces axes, il faut choisir des sous-espaces qui en réalisent un bon ajustement.

1.6 Objectifs exprimés en termes de liaisons (entre variables ou groupes de variables).

C'est en ces termes que l'analyse simultanée de plusieurs groupes de variables a été d'abord formulée. Nous faisons allusion ici au cas de deux tableaux

étudié par HOTELLING en 1936 [cf 12] à l'aide de ce qu'il appela l'analyse canonique. L'objectif général était d'étudier les relations entre deux ensembles de variables d'un point de vue synthétique, et non en considérant les relations entre les variables prises deux à deux. La formalisation mathématique de cet objectif est, dans l'analyse canonique : déterminer la combinaison linéaire des variables du premier groupe et la combinaison linéaire des variables du deuxième groupe les plus corrélées.

Cette idée est à la base de nombreuses généralisations au cas de J tableaux (J>2). L'objectif est alors de rechercher J combinaisons linéaires de variables (chaque combinaison est définie sur un groupe) telles que ces combinaisons soient les plus liées possibles. Selon les critères de liaison pour un ensemble de variables que l'on retiendra, on obtiendra des techniques distinctes (d'où les propositions de HORST [10], KETTENRING [13]..). Ces méthodes sont regroupées sous les termes d'analyse multicanonique, ou d'analyse canonique généralisée.

L'idée émise par CARROLL [5] est de chercher d'abord une variable générale la plus liée possible à tous les groupes de variables, puis, cette variable générale obtenue, de chercher dans chaque groupe la variable qui lui est la plus liée. Avec le critère qu'il a choisi, les calculs sont relativement simples. Mais, comme pour les autres analyses canoniques généralisées, et même l'analyse canonique, les difficultés d'interprétations font que ces techniques sont très peu utilisées. Elles le sont seulement dans le cas particulier des variables qualitatives. Pour deux variables qualitatives, l'analyse canonique se confond avec l'analyse des correspondances du tableau de contingence croisant ces deux variables ; pour un nombre supérieur à deux, l'analyse multicanonique au sens de Carrol se confond avec l'analyse des correspondances multiples [cf 8 et 19]. Mais en réalité, dans ces deux cas, c'est essentiellement avec l'optique d'analyse correspondances que ces techniques sont utilisées, c'est à dire avec les représentations des individus et des modalités des variables qualitatives (i.e. des centres de gravité des individus).

En plus des difficultés d'interprétation, un reproche que l'on fait à l'analyse canonique (généralisée ou non) est que la variance des groupes de variables expliquée par les variables canoniques obtenues peut être très faible. Pour augmenter la variance expliquée, des techniques un peu différentes de l'analyse canonique ont d'ailleurs été proposées [cf 21].

En résumé, le point de vue analyse des liaisons est intéressant, mais les résultats de l'analyse multicanonique ne sont vraiment

très utilisés que dans le cas de variables qualitatives.

Une direction de recherche peut être de tenter de généraliser l'analyse des correspondances multiples à des groupes de variables quelconques. C'est à dire de chercher des résultats admettant la double interprétation :

- en termes de liaison entre groupe.

- en termes de représentation des individus et des variables correspondant aux notions classiques en analyse factorielle.

Une démarche analogue à celle de CARROLL avec une mesure de liaison qui se confond avec la sienne dans le cas des variables qualitatives, nous assurera qu'il s'agit d'une généralisation de l'analyse des correspondances multiples. Dans le cas général, il serait souhaitable que cette mesure soit différente : d'une part pour augmenter la variance expliquée par les variables obtenues, et d'autre part pour obtenir des représentations des ensembles d'individus et de variables facilement interprétables (i.e. correspondant aux notions classiques en analyse en composantes principales).

1.7 Les modèles INDSCAL et IDIOSCAL

C'est une approche tout à fait différente puisqu'un modèle est proposé et qu'il s'agit d'en calculer les paramètres. Le modèle INDSCAL (Analysis of Individual Difference in Multidimensional Scaling) a été proposé essentiellement par CARROLL et CHANG [cf 4]. Il s'applique à des données plus générales que les nôtres ; matrices de distances entre objets ou matrices de similarités (pour nous, chaque groupe de variables définit une matrice de distances entre les individus).

Le modèle est le suivant : les distances entre individus peuvent se décomposer suivant un certain nombre de "facteurs" communs à tous les groupes, les poids affectés à chaque facteur différent suivant les groupes. Plus précisément notons :

- $F_s(i)$ la valeur du s-ième facteurs $(s = 1,S)$ pour l'individu i
- w_s^j le poids affecté à F_s par le j.ième groupe
- $d_j(i,i')$ la distance entre i et i' induite par le j.ième groupe

Le modèle s'écrit :

$$d_j^2(i,i') = \sum_{s=1}^{S} w_s^j \{F_s(i) - F_s(i')\}^2$$

Le modèle INDSCAL, lorsqu'il est adapté, permet de résumer d'une manière très synthétique les différences entre les nuages par les poids affectés à chaque "facteur".

Une manière équivalente d'exprimer le modèle est de dire que les individus peuvent être représentés pour tous les groupes par un même nuage de points dans un espace de dimension S ; les coordonnées d'un individu i sont égales à $F_s(i)$. Mais, à chaque groupe est associé sur cet espace un produit scalaire qui, exprimé dans la base canonique, est diagonal. Les éléments diagonaux sont les poids w_s^j.

Le modèle IDIOSCAL est plus général :les produits scalaires associés à chaque groupe sont quelconques.

Les paramètres à calculer sont les facteurs F_s et pour le modèle INDSCAL les poids w_s^j (pour IDIOSCAL les produits scalaires associés à chaque groupe).

Diverses solutions ont été proposées qui dépendent du type de critère de choix des paramètres que l'on cherche à optimiser. Citons pour INDSCAL la première technique proposée par CARROLL et CHANG [cf4] dont la convergence est assez lente et pour laquelle la solution risque d'être seulement un optimum local. Citons aussi SUMSCAL de DE LEEW et PRUZANSKY [cf16] beaucoup plus rapide.

Les facteurs communs permettent de donner une représentation "moyenne" des différents nuages. Les poids affectés aux facteurs par un groupe de variables donnent une représentation du nuage associé à ce groupe qui se déduit de la représentation moyenne par des homothéties axe par axe.

Il peut être intéressant de rapprocher cette représentation des nuages de celle qui a été discutée au paragraphe 1.4.1. Dans ce dernier nous cherchions des représentations qui soient des projections des nuages sur des espaces de petite dimension qui se ressemblent (dans le sens que les distances entre les points représentant le même individu soient assez faibles). Ici, avec INDSCAL on cherche des représentations telles que les facteurs, i.e. les projections des représentations sur les éléments de la base soient les mêmes à une homothétie près. Cette contrainte est très forte et ne permet pas d'obtenir pour les représentations des nuages des projections des nuages initiaux. (ceci ne serait possible que si le modèle était vérifié). Elle ne permet pas d'obtenir non plus des distances faibles entre les points représentant le même individu. Car ces distances peuvent être très grandes si les poids affectés à un même facteur ont des valeurs très éloignées.

Les critères de ressemblance entre les représentations sont donc de types très différents la contrainte pour INDSCAL est très forte. Les critères de qualité de représentation ne sont pas les mêmes non plus. Pour INDSCAL, le critère de qualité ne pourra être qu'un critère global sur les distances ou les produits scalaires, puisqu'on ne peut parler de pourcentage d'inertie extrait, par des représentations qui ne sont pas des projections des nuages.

Cependant, la représentation simultanée du paragraphe 1.4 a été aussi introduite en 1.4.2, à partir des préoccupations qui se rapprochent beaucoup d'INDSCAL. On y cherchait des "structures communes" aux nuages (i.e. des projections sur des

espaces de dimension 1 qui se ressemblent). On y cherchait aussi une structure moyenne pour représenter ces structures communes.

Cette structure moyenne pourrait fournir les facteurs du modèle INDSCAL .

Il ne resterait qu'à calculer les poids affectés par chaque groupe à ces facteurs pour avoir une solution pour INDSCAL. L'intérêt de cette solution, si elle est acceptable (ce dont il faudra discuter) sera

. de pouvoir se placer dans les cadres géométriques proposés pour résoudre les autres problèmes ;

. de juger de l'adéquation du modèle, non seulement avec un critère global qu'il sera nécessaire de calculer, mais aussi d'une manière très analytique puisqu'on disposera de deux représentations très proches des nuages, l'une avec les contraintes du modèle, l'autre sans ces dernières.

1-8 Conclusion

Nous avons dégagé très précisément un certain nombre d'objectifs pour comparer les groupes de variables :

. comparaison globale des groupes ;

. représentation simultanée des nuages d'individus définis par chaque groupe de variables (i.e. projections de ces nuages se ressemblant entre elles) et d'un nuage moyen ;

. comparaison des formes des nuages de variables (i.e. des composantes principales de chaque groupe) ;

. recherche de combinaisons linéaires des variables de chaque groupe corréllées entre elles (c.f. analyse multicanonique) ;

. modèle INDSCAL (facteurs communs aux groupes avec poids dépendant de chaque groupe).

2 - LES CADRES DE REFERENCES :

Ce paragraphe contient une étude détaillée des cadres de références et des objets placés dans ces cadres pour concevoir la méthode proposée.

Pour comparer des groupes de variables définies sur le même ensemble d'individus, on peut adopter deux points de vue :

a) considérer les variables en elles-mêmes,

b) s'intéresser à la structure définie sur l'ensemble des individus par le groupe de variables.

Nous verrons que ces deux points de vue aboutissent à introduire le même être mathématique pour caractériser un groupe de variables et à proposer la même méthode. Cette méthode donne plusieurs résultats se complètant ; certains permettent la comparaison directe des groupes de variables et d'autres la comparaison des structures définies sur l'ensemble des individus.

2 -1. L'espace R^I :

Pour étudier les variables elles-mêmes, le cadre classique est l'espace des fonctions numériques définies sur l'ensemble des individus I : l'espace R^I.

Chaque variable est représentée dans R^I par un vecteur.

Sur R^I, on définit une métrique euclidienne, notée D dont la matrice est diagonale et comprend sur la diagonale les poids des individus. On impose à ces poids d'être positifs et d'avoir une somme égale à un.

Rappelons que pour cette métrique, si v et w sont des variables quelconques, il y a les équivalences suivantes :

v *centrée* $\Leftrightarrow \langle v,u\rangle = v'Du = 0$ *avec* $u=(1,\ldots,1)$

v *centrée, réduite* $\Leftrightarrow \langle v,u\rangle = 0$ *et* $||v||=1$

corrélation $(v,w) = \cos(v,w) = \dfrac{\langle v\ ,\ w\rangle}{||v||\,||w||}$ *si* v *et* w *sont centrées*

Un groupe de variables est représenté par un groupe de vecteurs affectés chacun d'un poids.

2-2. les espaces R^{Kj} :

Nous nous intéressons maintenant aux structures induites sur I par les différents groupes de variables.
Ces structures sont obtenues à partir de distances euclidiennes. La distance induite par le j-ème groupe de variables est :

$$d^2(i,i') = \sum_{k \in K_j} m_k \{v_k(i) - v_k(i')\}^2$$

La représentation classique de I muni de ces distances est un nuage de points situé dans l'espace R^{Kj}. Les coordonnées de ces points sont contenues dans le tableau X_j. La métrique de R^{Kj} est une métrique diagonale des poids m_k des variables. On note M_j cette métrique.

Aux J groupes de variables étudiés, correspondent donc J nuages différents, chacun représentant I, situés dans des espaces distincts R^K_j. On note N^j_I le nuage associé au groupe K_j.

2-3. L'espace R^K

Considérons maintenant l'espace R^K, où K est l'union de tous les K_j. C'est l'espace naturellement associé à l'ensemble de toutes les variables sans distinction de groupes. La métrique diagonale des poids de ces variables est notée M.

Ce cadre est particulièrement intéressant car il permet de représenter simultanément les J nuages N^j_I et de construire un nuage moyen qui facilitera, comme nous l'avons déjà dit, les comparaisons de tous ces nuages.

2-3.1. Le nuage N_I associé à toutes les variables

L'espace R^K contient le nuage associé à l'ensemble K de toutes les variables dont les coordonnées sont contenues dans le tableau X. On note N_I ce nuage.

2-3.2. Représentation simultanée des J nuages N_I^j

Pour représenter simultanément ces J nuages dans l'espace R^K, il suffit de remarquer que R^K peut se décomposer en somme directe de sous espaces orthogonaux deux à deux et isomorphes aux espaces R^{K_j}.

$$R^K = \bigoplus R^{K_j}$$

Sur chacun de ces sous espaces, la métrique induite par M est la métrique M_j ; il s'agit donc d'un isomorphisme d'espaces euclidiens.

Le nuage N_I^j est donc représenté dans le sous espace R^{K_j} de R^K de la manière classique par ses coordonnées contenues dans le tableau que nous avons noté X_j. Ses coordonnées dans l'espace tout entier sont contenues dans un tableau noté $\tilde{X}_j$ de dimension I x K où X_j est complété par des zéros.

$$\tilde{X}_j = \begin{array}{c} \overbrace{}^{K} \\ \overbrace{}^{K_j} \end{array} \left\{ \begin{array}{|c|c|c|} \hline 0 & X_j & 0 \\ \hline \end{array} \right. I$$

Remarquons que le nuage N_I^j ainsi construit dans le sous-espace R^{K_j} de R^K est la projection orthogonale du nuage N_I sur R^{K_j}.

Les nuages N_I^j étant situés dans des sous-espaces orthogonaux, cette représentation simultanée est artificielle et inutilisable directement, mais elle va servir de base à une véritable représentation simultanée obtenue par projection sur des sous espaces de R^K.

2-3.3. Le nuage moyen N_I^*

Nous avons souligné déjà l'intérêt d'un nuage moyen pour faciliter les comparaisons. Différents critères peuvent être utilisés pour construire un nuage moyen. Celui qui s'introduit naturellement ici porte sur les carrés des distances.

En effet, dans le nuage N_I associé à l'ensemble de toutes les variables, le carré de la distance entre 2 points i et i' est la somme des carrés de leur distances dans les nuages N_I^j :

Notons : $d(i,i')$ la distance dans N_I
$d(i^j,i'^j)$ la distance dans N_I^j

Nous avons :

$$d^2(i,i') = \sum_{k\in K} m_k \, (v_k(i) - v_k(i'))^2$$

$$= \sum_{j\in J} \sum_{k\in K_j} (v_k(i) - v_k(i'))^2$$

$$= \sum_{j\in J} d^2_j(i,i')$$

Donc, à un facteur d'homothétie près, le carré de la distance dans N_I est la moyenne des carrés des distances dans les nuages N_I^j.

Nous considérons le nuage homothétique de N_I de rapport d'homothétie 1/J. Nous noterons N_I^* ce nuage et i^* le point de ce nuage représentant i. Ce nuage est intéressant : il a la même propriété de moyenne que N_I et en outre, chacun de ses éléments, i^* est situé au centre de gravité des points, i^j représentant i dans les nuages N_I^j

$$0\,i^* = \frac{1}{J} \sum_j 0\,i^j$$

Cette propriété se conservant par projection, si on projette simultanément dans R^K les nuages N_I^j et le nuage N_I^* sur des plans, la projection de i^* se trouvera au centre de gravité des points homologues des N_I^j ce qui facilitera beaucoup la comparaison.

2-3.4. Pondération des groupes de variables

Dans la construction du nuage moyen, se pose un problème d'équilibre de l'influence de tous les groupes de variables. En effet, si dans certains groupes, le nombre de variables ou leur poids est beaucoup plus important que dans d'autres, le nuage moyen reflétera essentiellement ces groupes en négligeant les autres. Pour éviter cela nous surpondérerons les variables en multipliant les poids de toutes les variables du groupe j par un même coefficient positif α_j. Le choix de ce coefficient fait partie intégrante de la méthode et sera discuté au § 3-5 en considérant tous les aspects du problème.

En multipliant les poids des variables par α_j, on peut effectivement modifier l'influence de chaque groupe de variables sur le nuage moyen puisque le carré de la distance entre deux éléments i et i' s'écrira :

$$d^2(i,i') = \sum_j \alpha_j \, d^2_j(i,i')$$

Lors de la recherche d'une représentation simultanée de ces nuages, on rencontre un autre aspect de ce problème.

En effet, une représentation simultanée des J nuages a pour but de comparer les structures induites sur I par les différents groupes de variables. Il est naturel de considérer que deux nuages homothétiques définissent des structures équivalentes. La comparaison des nuages par une représentation simultanée sera beaucoup plus simple si une normalisation préalable des nuages a rendu égaux deux nuages homothétiques.

Pour cela, il faudra transformer chaque nuage N_I^j par une certaine homothétie. Or, multiplier le poids des variables du groupe numéro j par un coefficient α_j revient exactement à appliquer au nuage N_I^j une homothétie de rapport $\sqrt{\alpha_j}$. Ainsi, la normalisation des nuages N_I^j conduit elle aussi à surpondérer chaque groupe de variables par un coefficient α_j.

par un coefficient α_j.

Afin de garder toute la cohérence des représentations des N_I^j et des nuages moyens N_I^*, (et les points de ce dernier au centre de gravité de leurs homologues dans les N_I^j), il faut que les coefficients α_j soient les mêmes pour la construction du nuage moyen et la normalisation des N_I^j.

Nous décrivons le choix des α_j au § 3-5. Souvent, dans la suite, lorsque cela ne sera pas précisé, ces coefficients α_j seront inclus dans le poids des variables. Ces poids se traduisent dans la métrique diagonale M de R^K.

2 -4. Dualité d'un nuage d'individus et d'un nuage de variables :

Nous étudions ici les relations entre un nuage de variables et le nuage d'individus qui lui est associé.

Pour simplifier les notations, nous parlons du nuage N_K et du nuage N_I, mais il est bien évident que les résultats s'appliquent à tous les couples de nuages N_I^j et N_K^j, en remplaçant X par X_j, R^K par $R^K j$ et M par M_j.

2-4.1. Projection des nuages sur un axe :

Rappel : Le nuage des individus N_I est situé dans l'espace R^K et le nuage des variables N_K, dans l'espace R^I. Le poids des variables dans le nuage N_K définit la métrique diagonale M de R^K et réciproquement le poids des individus dans N_I définit la métrique diagonale D de R^I. Ceci est indispensable pour que les résultats que nous indiquons ci-dessous soient vérifiés.

Soit a un vecteur normé de R^I. La projection du nuage N_K sur l'axe {a} définit une fonction numérique sur K, donc un élément de R^K. Cette fonction notée b s'écrit en fonction de a :

$$b = X' D a$$

On définit ainsi l'application X'D de R^I dans R^K.

De plus, la longueur de b nous renseigne sur l'inertie de cette projection car :

$$\text{Inertie de } N_K \text{ sur a} = (a' D X)\, M(X' D a)$$
$$= ||b||^2$$

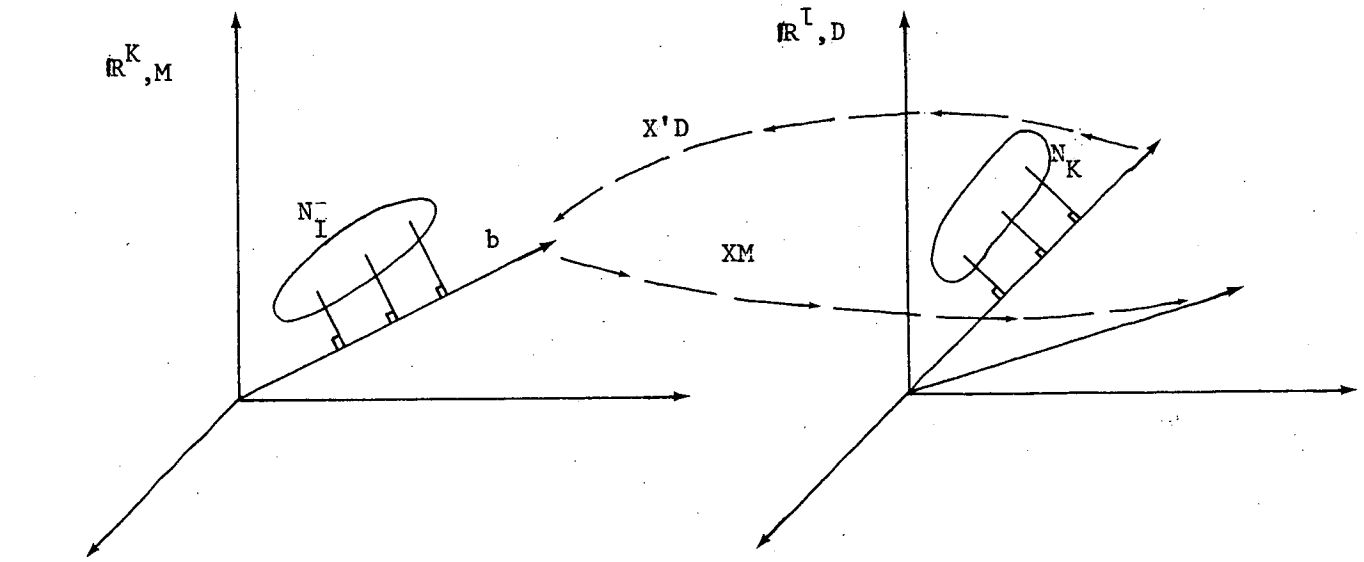
De même, la projection du nuage N_I définit l'application XM de R^K dans R^I. Si l'on prend le vecteur b ci-dessus, la projection c de N_I sur l'axe engendré par b s'écrit :

$$c = X M b / ||b||$$

L'inertie de cette projection est encore le carré de la norme de c.

$$\text{Inertie de } N_I \text{ sur l'axe } \{b\} = \left(\frac{b'}{||b||} M X'\right) D \left(X M \frac{b}{||b||}\right)$$

$$= ||c||^2$$



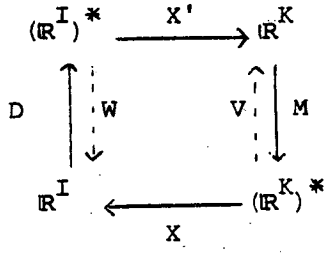
Projection de N_I sur b :XMb

projection de N_K sur a : X'Da

2-4.2. Schéma de dualité

Les métriques D et M définissent des isomorphismes de R^I et de R^K dans leurs duaux noté $(R^I)^*$ et $(R^K)^*$. Il est donc naturel de considérer l'application de $(R^K)^*$ dans R^I définie par X.

Le schéma de dualité [cf 3] résume l'ensemble de ces diverses applications :



V = X'DX

W = XMX'

Ce schéma est complété par les applications W = X'DX de $(R^I)^*$ dans R^I et V de $(R^K)^*$ dans R^K.

En composant les applications W et D, on obtient un opérateur de R^I, WD.

Pour chaque groupe de variables, on définit de même $W_j = X_j M_j X'_j$ et les opérateurs $W_j D$.

Cas des axes d'inertie :

Partant d'un vecteur a de R^I, on obtient donc, en appliquant X'D, la projection b de N_K sur {a}, puis en appliquant XM, la projection c de N_I sur b. Généralement, les vecteurs a et c ne sont pas colinéaires :

$$c = \frac{1}{||b||} X M X' D a = \frac{1}{||b||} W D a$$

Le vecteur c est alors la projection de N_I sur l'axe X' D a.

Ces vecteurs ne sont colinéaires que si a est vecteur propre de WD. C'est alors un axe d'inertie du nuage $N_{K'}$. Le vecteur b = X' D a est aussi vecteur propre de X' D X M et axe d'inertie de N_I associé au même moment d'inertie λ.

2-4.3. L'ellipsoide des projections d'un nuage :

Nous aurons besoin dans la suite de caractériser dans R^I les vecteurs représentant la projection du nuage N_I sur un axe. Nous avons vu que ce sont des vecteurs dont le carré de la norme est égal à l'inertie de cette projection.

Nous allons montrer que ces vecteurs sont contenus dans le sous-espace de R^I, noté E, engendré par l'ensemble des variables et qu'ils forment un ellipsoïde. L'équation de cet ellipsoïde est $c' W^{-1} c = 1$ ou W^{-1} est l'inverse de la restriction de W à E.

Démonstration :

Nous venons de voir que la projection c de N_I sur un axe {b} de R^K s'écrivait :

$$c = \frac{X M b}{||b||_M}$$

Supposons d'abord que b appartient au sous-espace engendré par N_I, il appartient à l'image de X'D et s'écrit en fonction d'un vecteur a de R^I :

$$b = X'Da$$

D'où c s'écrit :

$$c = \frac{X\ M\ X'\ D\ a}{||b||_M} \Rightarrow \frac{W\ D\ a}{\sqrt{a'\ D\ W\ D\ a}}$$

Or l'image de W est exactement le sous-espace E engendré par les variables. Donc c appartient à E.

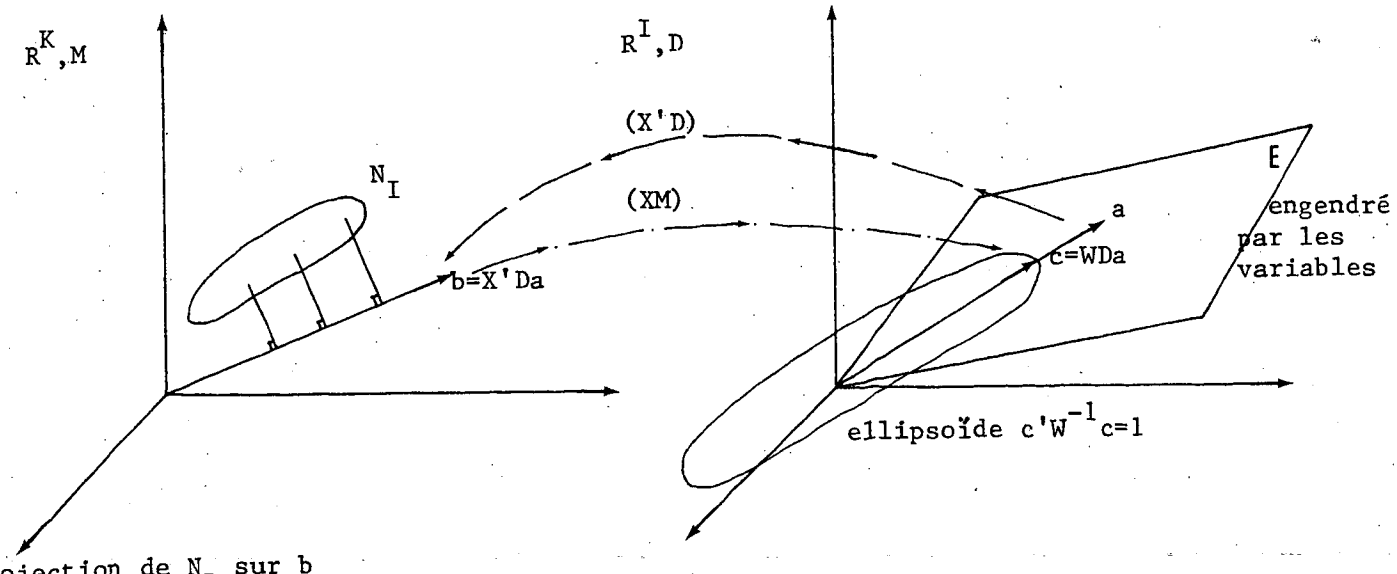
Définissons maintenant W^{-1}. Comme WD restreint à une application de E dans E est bijective et que D est bijective aussi, W est une bijection de D(E) dans E. On note W^{-1} l'inverse de cette bijection, qui est donc une application de E dans D(E), et qui définit une forme bilinéaire symétrique sur E.

Vérifions maintenant que $c'\ W^{-1} c = 1$ pour toute projection c de N^I. Puisque c appartient à E, W^{-1} est définie sur c. Elle s'écrit :

$$c'\ W^{-1} c = \frac{(a'D\ W)\ W^{-1}(W\ D\ a)}{a'\ D\ W\ D\ a}$$

$$= 1$$

Réciproquement, si c appartient à E, il est image par W D d'un vecteur a, et est colinéaire à la projection de N^I sur X' D a. La contrainte c'W c = 1 fixe sa longueur et implique l'égalité avec cette projection. c.q.f.d.



c=projection de N_I sur b

$||c||^2$ = inertie de cette projection

Voyons maintenant le cas où l'axe b sur lequel on projette N_I n'appartient pas au sous-espace engendré par N_I. Notons alors :

$\tilde{b}$ la projection de b sur ce sous-espace
θ l'angle entre b et $\tilde{b}$
$\tilde{c}$ la projection de N_I sur $\tilde{b}$
c la projection de N_I sur b

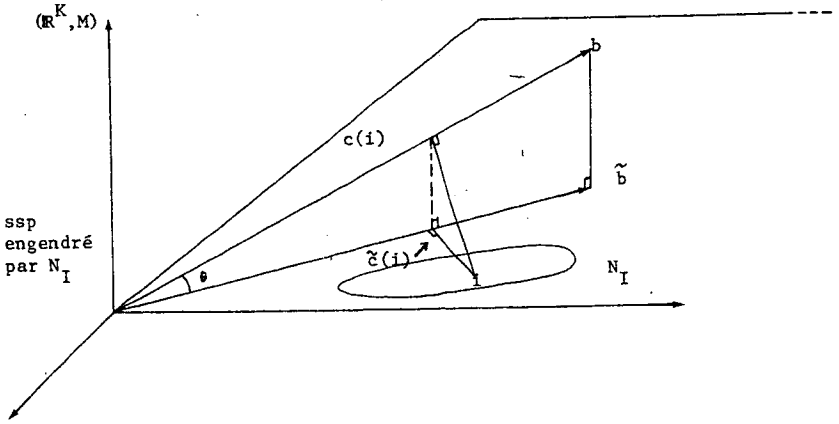
Nous allons montrer que :

$$c = \tilde{c} \cos \theta$$

d'où

$$c \in E \text{ et } c' W^{-1} c = \cos^2 \theta \leqslant 1$$

perpendiculaires, la projection d'un point i sur b est obtenue en projettant i d'abord sur le plan $|b,\tilde{b}|$, puis sur b. Il est facile de voir que la projection d'un point de N_I sur le plan $|b,\tilde{b}|$ est exactement la projection sur $\tilde{b}$. En ensuite sur b, on multiplie les longueurs des projections par cos θ.



Remarque : cas des axes d'inertie

La projection de N_I sur un axe étant un vecteur c de R^I tel que $c'W^{-1}c = 1$ et l'inertie de cette projection étant $||c||_D^2$, la projection d'inertie maximum est le vecteur de R^I rendant $||c||_D^2$ maximum sous la contrainte $c'W^{-1}c = 1$.
On retrouve les vecteurs propres de WD, i.e. axes d'inertie du nuage N_K qui sont confondus avec les axes principaux de l'ellipsoïde W^{-1}.

Pour les points P situés sur ces axes, on a :

$||\ \cdot\ ||^2$ = inertie de la projection de N_I
= inertie de la projection de N_K sur OP

Dans les autres directions, la deuxième ligne n'est pas vérifiée. Les vecteurs u de R^I dont le carré de la longueur est égal à l'inertie de la projection de N_K sur u ne forment d'ailleurs pas un ellipsoide. Par contre, les vecteurs v de l'ellipsoïde $v'DWDv$ ont pour norme carrée l'inverse de cette valeur.

2-5. L'espace $(R^I)^2$:

Pour comparer globalement les groupes de variables, on se place dans un espace euclidien de dimension $(\text{Card } I)^2$. Dans cet espace chaque groupe de variables est représenté par un vecteur unique.

Il existe plusieurs représentations isomorphes : dans un espace de matrices, dans le produit tensoriel $R^I \otimes R^I$, dans un espace d'opérateurs. Les points de vues offerts par ces différentes représentations serviront. Lorsque l'on ne précise pas le point de vue, on parle simplement de l'espace $(R^I)^2$.

Ces représentations possèdent deux propriétés intéressantes :

a) La connaissance du vecteur représentant un groupe de variables permet de décrire ce groupe : soit dans R^I, soit par les distances qu'il induit sur l'ensemble I.

b) La structure euclidienne de $(R^I)^2$ permet de comparer les groupes de variables . Nous vérifierons sur plusieurs cas particuliers que cette structure correspond bien à ce que l'on souhaite.

2-5.1 description d'un groupe de variables par W_j :

Nous montrons ici comment s'introduisent naturellement les vecteurs de $(R^I)^2$ pour décrire à la fois le groupe des variables dans R^I et les distances induites sur I par ce groupe.

2-5.1.a. W_j Matrice de produits scalaires :

Commençons par les distances induites sur I : ces distances peuvent être décrites dans des matrices symétriques de dimension Card I. Mais ces matrices sont peu maniables.

Or ces distances peuvent se déduire des produits scalaires des vecteurs joignant l'origine aux points représentant les individus dans $R^K j$. Ces produits scalaires s'écrivent aussi dans une matrice symétrique.

Mais, ces derniers dépendent de l'origine des axes, alors que les distances n'en dépendent pas. Pour qu'il y ait bijection entre matrice de produits scalaires et matrice de distances, il suffit de fixer l'origine des axes au centre de gravité du nuage.

Ceci revient à centrer les variables, et, sauf expression du contraire, nous considèrerons dans la suite que les variables sont centrées.

Dans ce cas, la matrice de produits scalaires est indépendante de la représentation euclidienne choisie pour le nuage d'individu. Rappelons que cette matrice qu'on note W se déduit des distances par la formule suivante :

$$W(i,i') = \frac{1}{2}(d^2_{i.} + d^2_{i'.} - d^2(i,i') - d^2_{..})$$

où $d^2_{i.} = \sum_{i' \in I} p_{i'} d^2(i,i')$ et $d^2_{..} = \sum_{i \in I} p_i d^2_i$.

Pour le groupe j, cette matrice s'écrit facilement en fonction de X_j, de sa transposée X'_j et de la matrice diagonale M_j. En effet, X_j contient les coordonnées des individus dans R^{K_j}. D'où :

$$W_j = X_j \, M_j \, X'_j$$

Les colonnes de la matrice X_j sont les coordonnées des variables v_k composant le groupe X_j et les éléments de la matrice diagonale M_j sont leurs poids m_k. La matrice W_j se décompose donc sous la forme :

$$W_j = \sum_{k \in K_j} m_k \, v_k v'_k$$

2-5:1.b. W tenseur d'inertie :

Replaçons nous dans l'espace R^I pour comparer directement les groupes de variables.

Dans cet espace, chaque groupe est représenté par un ensemble de vecteurs muni de poids. Pour comparer ces groupes, on peut comparer les positions et les formes de ces nuages.

Or, il existe une notion qui caractérise la forme d'un nuage, c'est sa forme quadratique d'inertie, ou tenseur d'inertie.

En notation tensorielle, il s'écrit :

$$W_j = \sum_{k \in K_j} m_k v_k \otimes v_k$$

Soit, en notation matricielle :

$$W_j = \sum_k m_k \, v_k \, v_k'$$

On retrouve le W_j précédent, d'où l'identité des notations.

Rappelons que ce tenseur caractérise la forme du nuage par son inertie. En effet, il permet - avec la métrique D de l'espace R dont il est indépendant - de calculer :

a) l'inertie de la projection du nuage dans toutes les directions de l'espace,

b) les moments d'inertie et les directions des axes d'inertie du nuage.

Un tenseur peut être considéré soit comme un élément du produit tensoriel $R^I \otimes R^I$; soit comme une forme bilinéaire symétrique sur le dual de R^I noté $(R^I)^*$; soit comme une application linéaire de $(R^I)^*$ dans R^I.

Pour calculer l'inertie de la projection du nuage sur une direction u de R^I, on applique la forme bilinéaire symétrique à D u où D est considérée. comme une application linéaire de R^I dans son dual :

$$W_j(D\,u, D\,u) = \text{Inertie de la projection du nuage sur } u$$
$$= \sum_{k \in K_j} m_k \, (v_k' \, D\, u)^2$$
$$= u' \, D \, W_j \, D \, u$$

Pour calculer les axes d'inertie et les moments d'inertie du nuage, on diagonalise l'opérateur $W_j D_{p_I}$ où W_j est maintenant considéré comme une application linéaire de $(R^I)^*$ dans R^I.

2-5.1.c. Comparaison entre W_j et le sous espace engendré par un groupe de variables pour caractériser ce groupe.

Pour caractériser un groupe de variables, nous utiliserons le tenseur W_j, ou - ce qui est équivalent en un certain sens, comme nous le verrons - l'opérateur $W_j D$.

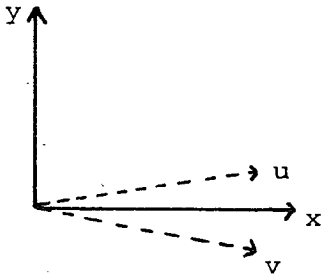
En analyse canonique, multicanonique, en regression [cf.4] on associe à chaque groupe de variables, le sous-espace de R^I qu'elles engendrent. On caractérise ce sous-espace, par son opérateur de projection orthogonale.

Il s'agit là d'une différence essentielle entre les méthodes. La comparaison entre ces deux approches lient en trois points :

a) W_j permet de calculer les distances entre les individus de I définies par le groupe de variables. Le projecteur ne le permet pas.

b) W_j contient plus d'informations que le projecteur puisque l'image de W_j est exactement le sous-espace engendré par les variables. Il en est de même pour $W_j D$, puisque D est un isomorphisme. La différence entre $W_j D$ et l'opérateur de projection orthogonale provient seulement de leurs valeurs propres non nulles. Elles sont toutes égales à 1 pour le projecteur et généralement différentes entre elles pour $W_j D$.

c) W_j est beaucoup plus stable que l'opérateur de projection orthogonale. L'exemple ci-dessous le montre.



Le groupe de variables u,v engendre le plan (x,y). Le tenseur donne une grande importance à la direction x de grande inertie et une très faible à y, alors que le projecteur donne la même importance à ces deux directions. Une petite variation de u conservera à peu près la direction x mais peut modifier totalement l'autre direction du plan, le projecteur sera très différent alors que le tenseur variera peu.

2-5.1.d. Cas des variables qualitatives :

Une variable qualitative définit une partition de l'ensemble des individus. Si on associe à cette variable l'ensemble des variables indicatrices des classes de cette partition, on peut dire qu'une variable qualitative est un cas particulier de groupe de variables.

Ce cas particulier est intéressant puisque l'on sait très bien le traiter par l'analyse des correspondances multiples en analysant le

tableau disjonctif complet dont les colonnes sont justement les variables indicatrices évoquées ci-dessus.

Pour décrire une variable qualitative, c'est le sous-espace engendré par les variables indicatrices qui est intéressant et non un quelconque tenseur d'inertie. En effet, ce sous-espace est celui des fonctions prenant la même valeur pour des individus appartenant à la même classe de la partition, il décrit donc parfaitement cette partition. [cf 8]

Les variables indicatrices étant orthogonales deux à deux, la dimension de ce sous-espace est égale au nombre de modalités de la variable.

La somme de ces variables est toujours égale au vecteur $u = (1,\ldots,1)$. Le sous-espace E qu'elles engendrent peut se décomposer en deux sous-espaces orthogonaux, l'un de dimension 1 engendré par u, et l'autre E de dimension égale au nombre de modalités moins 1. Ce dernier est tout simplement le sous-espace engendré par les variables indicatrices centrées, il caractérise donc aussi la variable qualitative. On considérera l'un ou l'autre de ces deux sous-espaces.

Un sous-espace est caractérisé par l'opérateur de projection orthogonale sur lui-même. . Pour faire entrer les variables qualitatives dans le cadre général, il suffit que l'opérateur $W_j D$ soit égal à ce projecteur i.e. que tous les moments d'inertie du nuage des indicatrices soient égaux à 1.

Remarquons que ceci est réalisé en analyse des correspondances multiples : la variable indicatrice k_{ij} est remplacée par son profil k_{ij}/k_j et on lui affecte le poids k_j ce qui lui donne une inertie égale à un. Ces variables étant orthogonales deux à deux, tous les moments d'inertie sont égaux.

Dans la suite, nous appellerons variable qualitative un groupe de variables indicatrices (ou de profils) pondérées pour rendre égaux les moments d'inertie.

2-5.2. Définition de la structure euclidienne de $(R^I)^2$:

Dans le paragraphe précédent, nous avons associé à chaque groupe de variables un vecteur W_j de l'espace $(R^I)^2$. Pour comparer ces W_j entre eux, il reste à définir sur cet espace une structure euclidienne, qui est la notion de proximité la plus facile à manier.

Nous la définissons d'abord en considérant $(R^I)^2$ comme le produit tensoriel, $R^I \otimes R^I$ car la métrique D de R^I induit sur cet espace une métrique notée classiquement $D \otimes D$. Nous le traduisons ensuite dans les espaces isomorphes : l'espace des matrices et l'espace des opérateurs de R^I.

2-5.2.a. $(R^{I2}) = R^I \otimes R^I$

Rappelons d'abord que les éléments de $R^I \otimes R^I$ sont des combinaisons linéaires de tenseurs particuliers, appelés tenseurs de rang 1 . Ces tenseurs s'écrivent $u \otimes v$ où u et v sont des vecteurs quelconques de R^I.

Le rang d'un tenseur est le nombre minimum de tenseurs de rang 1 permettant de le décomposer.

Rappelons aussi que si u s'écrit : $u = \Sigma \alpha_i e_i$, alors $u \otimes v$ se décompose : $u \otimes v = \sum_i \alpha_i (e_i \otimes v)$. Il en est de même pour v.

Si les e_i forment une base de R^I, les $e_i \otimes e_i$, forment une base de $R^I \otimes R^I$. Les coordonnées de W_j dans cette base sont dans la matrice associée à W_j.

Par définition, le produit scalaire entre deux tenseurs de rang 1, $u \otimes v$ et $x \otimes y$ vaut :

$$\langle u \otimes v, x \otimes y \rangle_{D \otimes D} = \langle u, x \rangle_D \ \langle v, y \rangle_D$$

$$= u' \ D \ x \ v' \ D \ y$$

Sa valeur pour un tenseur de rang quelconque s'en déduit par bilinéarité :

calculons-la pour deux tenseurs W_j et $W_{j'}$, associés aux groupes de variables K_j et $K_{j'}$, affectées des poids m_k et n_ℓ :

$$\langle W_j, W_{j'} \rangle = \Big\langle \sum_{k \in K_j} m_k \, v_k \otimes v_k \ , \ \sum_{\ell \in K_{j'}} n_\ell \, x_\ell \otimes x_\ell \Big\rangle$$

$$= \sum_{k,\ell} m_k \, n_\ell \ (v'_k \ D \ x_\ell) \ (v'_k \ D \ x_\ell)$$

$$= \text{trace} \ (W_j \ D \ . \ W_{j'} \ D)$$

2-5.2.b. Tenseurs W_j et opérateurs $W_j D$

On reconnait dans le deuxième terme le produit scalaire classique entre les opérateurs D symétriques $W_j D$ et $W_{j'} D$. Il est donc équivalent de travailler avec W_j dans l'espace des tenseurs ou avec $W_j D$ dans l'espace des opérateurs muni de ce produit scalaire. Suivant les cas, on choisit la notion la plus facile à manier. Cette équivalence justifie la pratique maintenant classique des opérateurs $W_j D$ avec ce produit scalaire. [cf. 6 et 3]

2-5.2.c. Distance entre matrices de produits scalaires

Notons $a_{ii'}$, le terme général de la matrice W_j et $b_{ii'}$, celui de la matrice $W_{j'}$ et p_i le poids de l'individu i. On a :

$$D^2(W_j, W_{j'}) = \sum_{i,i'} (a_{ii'} - b_{ii'})^2 \, p_i \, p_{i'}$$

C'est une différence terme à terme pondérée par les poids des individus.

2-5.3. Interprétation du produit scalaire. Liaison entre deux groupes de variables

Le produit scalaire que nous avons défini sur les tenseurs ou opérateurs associés à un groupe de variables nous servira de mesure de liaison entre ces groupes.

Nous allons étudier ici cette mesure en prenant quelques cas particuliers, groupes de variables réduits à un élément, variables qualitatives, etc... Nous montrerons qu'elle correspond à ce que l'on peut attendre d'une mesure de liaison entre deux groupes de variables.

Remarquons d'abord que si $W_j = W_{j'}$, les distances sur I induites par les deux groupes sont les mêmes, les nuages sont identiques. Si $W_j = \lambda W_{j'}$, les deux nuages sont homothétiques, de rapport d'homothètie λ. Il parait souvent logique de considérer alors les deux groupes comme analogues. On retrouve le problème de normalisation des W_j déjà abordé au § 2-3.3 dont nous discuterons plus tard (cf § 3-5).

En réalité, c'est surtout la direction des vecteurs W_j qui est importante, c'est donc les produits scalaires ou les cosinus des angles que nous manierons plutôt que les distances entre les points.

2-5.3.a. Les deux groupes ont une seule variable

C'est le cas le plus simple. Dans ce cas la mesure de liaison habituelle est le coefficient de corrélation linéaire.

Considérons deux variables centrées réduites de poids 1 et notons u et v ces deux variables.

Le tenseur associé au premier groupe, réduit à u s'écrit :

$$W_1 = u \otimes u$$

et

$$||W_1||^2 = \langle W_1, W_1 \rangle$$

$$= \langle u,u \rangle_D \langle u,u \rangle_D$$

$$= 1$$

Le second $W_2 = v \otimes v$ est aussi de norme 1.

Leur produit scalaire est :

$$\langle W_1, W_2 \rangle = \langle u,v \rangle_D \langle u,v \rangle_D$$

$$= \cos^2 \Theta$$

où Θ est l'angle entre u et v.

Ce produit scalaire s'interprête comme le carré du coefficient de corrélation entre u et
Il ressemble donc bien évidemment au coefficient de corrélation.
Sur le passage de ce coefficient à son carré, on peut faire plusieurs remarques :

La première est que le produit scalaire entre les tenseurs est toujours positif, contrairement au coefficient de corrélation. Le sens des variables n'influe pas sur les distances induites, et des liaisons négatives entre groupes de variables n'ont pas de sens.

La seconde est que la liaison entre les tenseurs est plus petite que celle qui existe entre les variables, ce que nous retrouverons plus tard.

Si u est muni du poids m_1 et v du poids m_2 , la liaison entre u et v s'écrira : $m_1\ m_2\ \cos^2\theta$.

2-5.3.b. Groupe d'une variable et groupe quelconque

Soit u, la variable du premier groupe que nous supposerons réduite et de poids 1.

Soient v_k , $k \in K_2$ les variables du second groupe et m_k leur poids.

Par définition, le produit scalaire entre les tenseurs associés W_1 et W_2 vaut :

$$\langle W_1, W_2 \rangle = \langle u \otimes u, \sum_k m_k\ v_k \otimes v_k \rangle$$

$$= \sum_k m_k\ (\langle u, v_k \rangle)^2$$

$$= \sum_k \text{Inertie de la projection de } v_k \text{ sur } u$$

La seconde égalité montre que la liaison entre u et le groupe des v est la somme des liaisons entre u et chacune des v_k .

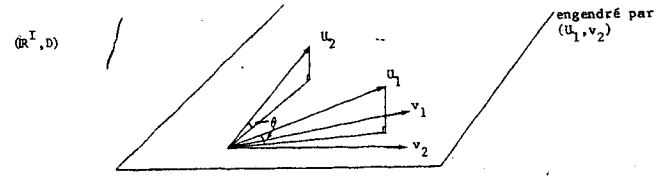
La troisième égalité conduit à un moyen de calcul simple de ce produit scalaire. Notons u' la projection de u sur le sous espace E engendré par les v_k, et θ l'angle entre u et u'. On montre facilement que ce produit scalaire est encore égal au produit par $\cos^2\theta$ de la somme des inerties des projections des v_k sur u' :

$$\langle W_1, W_2 \rangle = \cos^2\theta \sum_k (\text{Inertie des } v_k \text{ sur } u')$$

Cette dernière remarque permet de comparer facilement cette mesure de liaison au coefficient de corrélation multiple entre u et les v_k. En effet, lorsque les variables sont centrées, ce que nous supposerons ici, ce coefficient de corrélation multiple est $\cos\theta$.

Si l'inertie du nuage des v_k dans la direction u' vaut 1, le produit scalaire vaut $\cos^2\theta$. Ceci est réalisé pour une variable qualitative puisque nous imposons au nuage des indicatrices d'avoir des moments d'inertie tous égaux à 1.

Sinon, à angle égal entre u et le sous espace engendré par les v_k, le produit scalaire est grand si u se projette sur une direction de grande inertie et petit s'il se projette sur une direction de faible inertie. Le graphique ci-dessous l'illustre :



Si u se projette suivant la direction x, la liaison est forte et le produit scalaire grand, s'il se projette suivant la direction y, la liaison est faible et le produit scalaire petit.

- u_1 et u_2 ont le même coefficient de corrélation multiple avec le groupe (v_1 et v_2)
- mais u_1 est plus lié que u_2 à la direction d'inertie maximum de (u_1, v_2), ce que traduit la mesure de liaison proposée.

On retrouve ici les conclusions du paragraphe II-3.1.c. : dans le cas général, le produit scalaire entre tenseurs contient plus d'information , est plus stable et correspond mieux aux buts poursuivis que les notions attachées aux sous-espaces engendrés par les variables.

Remarquons que ce produit scalaire atteint sa valeur maximum lorsque $\cos^2 \theta$ vaut 1 et que l'inertie des projections des v_k est maximum. La variable u est alors dans la direction du premier axe d'inertie du nuage des v_k et le produit scalaire est égal au plus grand moment d'inertie de ce nuage, i.e. à la plus grande valeur propre de W_2 D. Sa valeur minimum, zéro, est atteinte lorsque u est orthogonal à E, i.e. non correlé aux v_k.

2-5.3.c. Deux groupes quelconques

Le produit scalaire entre les tenseurs W_j et $W_{j'}$, associés aux groupes de variables notées dans les sous-tableaux X_j et $X_{j'}$, s'écrit en fonction des covariances entre les variables des deux groupes.

Ecrivons ce produit scalaire :

$$\langle W_j, W_{j'} \rangle = \left\langle \sum_{k \in K_j} m_k \, v_k \times v_k \, , \sum_{\ell \in K_{j'}} n_\ell \, v_\ell \times v_\ell \right\rangle$$

$$= \sum_{k,\ell} (v'_k \, D \, v_\ell)^2 \, m_k \, n_\ell$$

Il peut s'interpréter comme la somme pondérée par les poids des variables des

carrés des covariances entre les variables des deux groupes. Ce produit scalaire est toujours positif ou nul ; il est nul lorsque les variables des deux groupes sont non correlées

On peut écrire ce produit scalaire sous forme condensée avec les matrices de covariance entre les variables des deux groupes. Notons $C_{jj'}$ et $C_{j'j}$ ces deux matrices transposées l'une de l'autre :

$$\begin{aligned} \langle W_j, W_{j'} \rangle &= \text{Trace } (W_j D)(W_{j'} D) \\ &= \text{Trace } X_j M_j (X'_j D X_{j'}) M_{j'} X'_{j'} D \\ &= \text{Trace } (X'_{j'} D X_j) M_j (X'_j D X_{j'}) M_{j'} \\ &= \text{Trace } C_{j'j} M_j C_{jj'} M_{j'} \end{aligned}$$

Remarquons que $||W_j||^2 = \text{Trace } (W_j D)^2$

$= \Sigma$ (valeurs propres de $W_j D)^2$

Si les opérateurs $W_j D$ et $W_{j'} D$ sont des opérateurs de projections orthogonales -ce qui est le cas pour les variables qualitatives- ce produit scalaire est la somme des carrés des corrélations canoniques.

Remarquons enfin que ce produit scalaire a une valeur beaucoup plus élevée lorsque ce sont des directions de grande inertie des deux groupes qui sont proches, que lorsque ce sont des directions de faible inertie.

2-5.3.d. Cas des variables qualitatives

La liaison entre 2 variables qualitatives est décrite entièrement par le tableau de contingence croisant les deux partitions. Elle est classiquement mesurée par le Φ^2 qui vaut, en notant k_{ij} le terme général du tableau de contingence, k_i la somme d'une ligne, k_j celle d'une colonne et n le cardinal de l'ensemble des individus :

$$\Phi^2 = \frac{1}{n} \sum_{i,j} \frac{(k_{ij} - k_i k_j)^2}{k_i k_j}$$

Le Φ^2 est nul si les variables sont indépendantes et croît avec leur dépendance.

Un calcul simple [cf. 3] montre que le produit scalaire entre les tenseurs vaut Φ^2+1 si on considère le sous-espace des variables indicatrices et Φ^2 si l'on considère les variables centrées.

2-5.4. Le tenseur du nuage moyen

Le nuage moyen introduit dans R^k était le nuage associé à l'ensemble de toutes les variables, chaque groupe de variables étant éventuellement surpondéré par un coefficient α_j.

Il est facile de voir que le tenseur d'inertie W de toutes les variables s'écrit :

$$W = \alpha_1 W_1 + \ldots\ldots + \alpha_J W_J$$

Et réciproquement, toute combinaison linéaire positive des W_j est un tenseur d'inertie de l'ensemble des groupes de variables surpondérés. Les tenseurs forment donc un demi cône convexe de $(R^I)^2$.

Sous cette forme le problème du choix de α_j , qui détermine l'importance de chaque groupe de variables dans les distances entre points du nuage moyen, rejoint le problème d'une "normalisation" des W_j.

2-6. L'espace $(R^I)^*$

La définition de W_j comme forme bilinéaire symétrique sur $(R^I)^*$ permet de proposer une représentation du nuage des individus dans $(R^I)^*$ absolument équivalente à la représentation classique dans R^{K_j}, puisque les distances entre individus sont les mêmes.

En effet, W_j en tant que forme bilinéaire symétrique définit une métrique sur $(R^I)^*$. En réalité, W_j étant seulement semi définie, c'est un "semi" produit scalaire. Or, nous avons vu qu'il suffisait que les produits scalaires entre les individus soient donnés par W_j pour que les distances interindividuelles soient les distances voulues. En considérant la métrique W_j sur $(R^I)^*$, il suffit de placer les individus aux vecteurs de la base canonique (e_i) . Le vecteur e_i qui représente l'individu i est la forme linéaire

sur R^I qui, à toute fonction numérique sur I, associe la valeur de cette fonction pour l'individu i.

Dans cette nouvelle représentation des nuages de points associés aux groupes de variables, tous les nuages sont non seulement situés dans le même espace mais confondus. C'est la métrique de l'espace qui varie avec le groupe; soit W_j pour le nuage associé au groupe j et $\sum_j \alpha_j W_j$ pour le nuage moyen.

Les représentations du même nuage N_I^j dans $(R^I)^*$ et dans R^{K_j} sont liées. L'application X'_j de $(R^I)^*$ dans R^{K_j} applique le premier sur le second. Cette application conserve la métrique.

Soient $u, v \in (R^I)^*$

$$\langle u,v \rangle_{W_j} = u' W_j v$$

$$= u' X_j M_j X'_j v$$

$$= \langle X'_j u, X' v \rangle_{M_j}$$

Si u est un vecteur de $(R^I)^*$, la projection de N_I^j sur u est égale à la projection dans R^K de N_I^j sur $X'_j u$.

$$\text{Projection de } N_I^j \text{ sur } u = W_j u$$

$$= X_j M_j X'_j u$$

$$= \text{Projection de } N_I^j \text{ sur } X'_j u$$

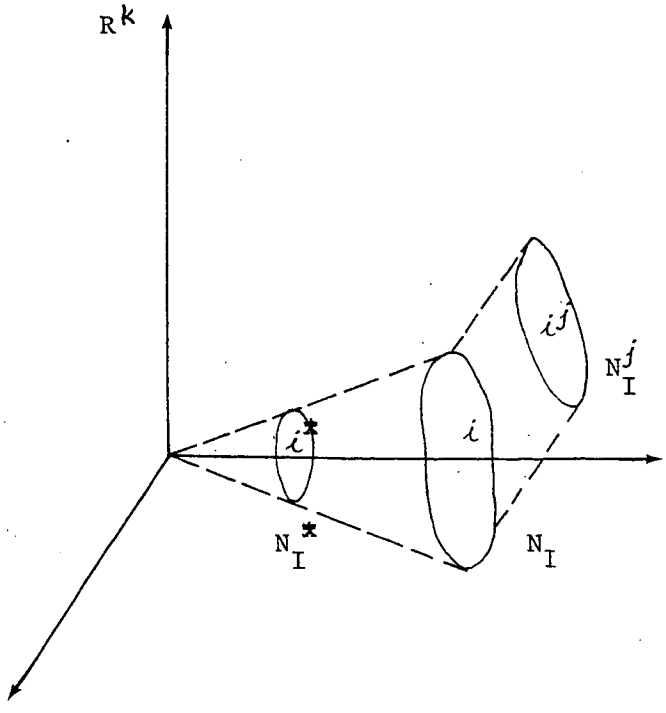
3. PROPOSITION D'UNE NOUVELLE METHODE

La principale originalité de la méthode proposée est qu'elle s'interprète simultanément dans tous les cadres de référence étudiés lors de la deuxième partie : R^K, R^I, R^{I2}, R^{I*}. Cette propriété est essentielle et justifie a posteriori l'ensemble des choix que l'on a pu faire dans sa construction. C'est pourquoi la méthode sera présentée successivement à l'intérieur de chacun des cadres de référence .

Cette présentation sera complétée par des paragraphes consacrés aux points suivants : discussion du choix des pondérations des groupes de variables, construction d'aides à l'interprétation, présentation résumée des principaux résultats, . prise en compte d'éléments supplémentaires.

3.1 Présentation dans R^K

3.1-1 Rappel des notations (et quelques nouvelles)



— $R^K = \oplus R^{K_j}$

i = point représentant l'individu $n°\,i$ du point de vue des K variables.

i^j = point représentant l'individu $n°\,i$ du point de vue des K_j variables du groupe j. $i^j \in R^{K_j}$

N_I = ensemble des points i.

N_I^j = ensemble des points i^j.

i^* = point homothétique de i, de rapport $1/J$. i^* est au centre de gravit de $\{i^j / i=1,I\}$

N_I^* = ensemble des points i^*.

.../...

3.1-2 Le problème de la représentation simultanée des intra-structures

Le cadre R^K présente l'avantage de contenir l'ensemble des nuages N_I^j. Il contient en outre le nuage N_I dont chaque point i est la somme de ses homologues i^j pour l'ensemble des nuages N_I^j. Il est donc bien adapté à la discussion de la représentation simultanée.

Ainsi que nous l'avons déjà indiqué, une représentation simultanée des nuages N_I^j est nécessairement artificielle puisque ces nuages n'évoluent pas dans le même espace. Néanmoins, si elle possède certaines propriétés, elle peut rendre de grands services dans l'interprétation. Pour notre problème, une représentation simultanée utile devra posséder les propriétés suivantes, déjà citées :

(p1) chacun des nuages N_I^j est bien représenté.

(p2) les points homologues i^j $(j=1,J)$ sont le plus rapprochés possible.

(p3) la représentation inclut un nuage compromis (=nuage moyen).

Nous ajouterons ici une autre propriété, qui se situe sur un autre plan, et dont la présence n'est pas intrinsèque à l'idée de représentation simultanée.

(p4) le repère de la représentation est interprétable directement à l'aide des variables.

Chacune de ces propriétés appelle quelques remarques.

Propriété p1

Soit N_I un nuage et N_I' sa représentation. Soit i un point de N_I et i' le point homologue de i dans N_I'. Le nuage N_I est bien représenté (par N_I') si les formes (id l'ensemble des distances inter-individuelles) se ressemblent. Le critère classique, pour apprécier cette ressemblance, est une somme de carrés des différences de distances du type :

$$\sum_{i\in I} \sum_{k\in I} (d(i,k) - d(i',k'))^2 \qquad \text{(à rendre petit)}$$

.../...

En analyse factorielle, on passe de N_I à N'_I par une projection. En ce cas, puisque $d(i',k') \leqslant d(i,k) \quad \forall i,k \in I$, le critère devient :

$$\sum_{i \in J} \sum_{k \in I} d(i',k')^2 \qquad \text{(à rendre grand)}$$

Ces considérations, appliquées à notre cas particulier où les nuages N_I^j évoluent dans R^K nous indiquent que, si nous décidons (ce que nous ferons) d'obtenir une représentation simultanée au moyen d'une projection des nuages N_I^j, la propriété (p1) peut s'exprimer ainsi :

p'1 : chacune des représentations des nuages est très dispersée.

Propriété p2

On retrouve ici une formulation d'un problème connu sous le nom d'analyse procustéenne, dont l'objectif est de comparer plusieurs configurations de points homologues. Pour cela, on essaie de superposer les configurations de façon à ce que les distances entre points homologues expriment une différence de forme entre les nuages et non l'hétérogénéité des repères qui les définissent.

Il est clair que cette propriété n'a de sens que si (p1) est réalisée. (Sinon, elle conduit à des représentations absurdes du type : tous les points de tous les groupes sont confondus). C'est en ce sens que doit être compris "le plus rapprochés possible" ; c'est-à-dire, "compte tenu d'une bonne représentation de chacun des nuages". Ainsi, dans le cas où les groupes de variables induisent des structures très différentes sur les individus, les points homologues dans la représentation simultanée seront peu rapprochés (ce qui ne les empêche pas de l'être le plus possible).

Il apparaît ainsi que les propriétés (p1) et (p2) sont concurrentes. En éliminant le cas limite (et facile à traiter sur le plan pratique) où les structures induites sont identiques, une représentation simultanée ne peut pas optimiser à la

.../...

fois la qualité de représentation des N_I^j et la proximité des points homologues i^j. Une représentation simultanée réalise donc toujours, implicitement ou explicitement, un compromis entre ces deux aspects.

Propriété p3

Il est toujours possible, une fois la représentation simultanée obtenue, d'y ajouter les points moyens des points homologues. Mais, pour être véritablement intéressante, cette configuration moyenne doit être l'image du nuage N_I ou N_I^* par la même application qui permet de passer des nuages N_I^j à leur représentation. Cette propriété sera réalisée si l'application précitée est une projection : la configuration moyenne est alors l'image de N_I^*.

Propriété p4

Le fait de pouvoir situer les axes de la représentation dans R^K permet de les interpréter comme combinaisons linéaires des variables initiales.

3.1-3 Le principe de la méthode

Le cadre R^K est un repère bien adapté à la recherche d'une représentation simultanée. L'analyse factorielle suggère de rechercher dans cet espace des axes sur lesquels projeter les nuages N_I^j ; $j=1,J$ et N_I^*. En procédant ainsi, la représentation obtenue possède automatiquement les propriétés (*p3*) et (*p4*).

Le problème se ramène donc au choix d'un compromis entre les propriétés (*p1*) et (*p2*). Pour cela, il convient d'abord d'expliciter les critères mis en jeu par ces deux propriétés.

.../...

Critère mis en jeu par (p1)

Les nuages N_I^j sont supposés centrés : tous ont leur centre de gravité en O.

La qualité de la représentation d'un nuage N_I^j par sa projection sur un axe de vecteur unitaire u est classiquement mesurée par l'inertie du nuage projeté que nous notons :

$I_0^u(N_I^j)$ (inertie du nuage N_I^j, par rapport à O, le long de la direction u)

Si l'on s'intéresse à l'ensemble des nuages N_I^j $j=1,J$, ce critère devient :

$$C1 = \sum_{j=1}^{j=J} I_0^u(N_I^j)$$

Cette inertie est celle de l'ensemble $\bigcup_j N_I^j$ des nuages N_I^j.

Remarquons que ce critère suppose implicitement une pondération entre les nuages. En particulier, il n'est pas tenu compte ici du fait que les nuages N_I^j peuvent avoir des inerties très différentes. Nous aborderons ultérieurement (§ 3.5) cet aspect de la discussion du choix d'une métrique dans R^K.

En se limitant à (p1) on serait conduit à chercher u tel que C1 maximum.

Critère mis en jeu par p2

La proximité entre les points de l'ensemble $\{i^j / j=1,J\}$ peut être mesurée par l'inertie du nuage associé à cet ensemble. L'inertie de ce nuage est calculée par rapport à son centre de gravité, soit i^* ($\in N_I$). En projection sur u, cette inertie sera notée :

- pour un point :

  $I_{i^*}^u(i^j)$ inertie du point i^j par rapport à i^* le long de la direction u.

- pour le nuage :

$$\sum_{j=1}^{j=J} I_{i^*}^u(i^j)$$

.../...

Comme l'on s'intéresse à l'ensemble des individus i, on considérera

$$C2 = \sum_{i=1}^{i=I} \sum_{j=1}^{j=J} I^{u}_{i^*}(i^j)$$

En appelant P la partition de $\bigcup_j N_I^j$ dont les classes sont les ensembles $\{i^j \mid j=1,J\}$ d'éléments homologues, le critère C2 s'interprète comme une inertie intra-classe.

Plus encore que le critère C1, les pondérations implicites des nuages N_I^j sous-jacentes à l'utilisation de ce critère n'apparaissent pas au premier abord. Elles interviennent en fait dans $\sum_{j=1}^{j=J} I^{u}_{i^*}(i^j)$ et appellent la même remarque.

En se liminant à (p2), on serait conduit à chercher u tel que C2 minimum.

Compromis entre les deux critères

En considérant le nuage $\bigcup_j N_I^j$ et la partition P précédemment définie, nous cherchons un axe de vecteur unitaire u.

1) d'inertie totale élevée (critère C1)

2) d'inertie intra-classe petite (critère C2).

Un compromis possible consiste à chercher u tel que le rapport C1/C2 soit maximum. On aboutit alors à une formulation analogue à celle de l'analyse discriminante, dont la résolution suppose une égalité des tenseurs intra-classes, (attention, les classes sont formées ici de points homologues). Nous n'avons pas exploré cette voie pour l'instant.

L'autre compromis consiste à chercher u tel que C1-C2 soit maximum. C'est celui que nous avons retenu.

3.1-4 *Interprétation en termes d'analyse factorielle*

En appliquant le théorème de Huyghens à l'ensemble $\bigcup_j N_I^j$ muni de la partition P, on obtient :

I totale = I inter-classes + I intra-classe

.../...

Soit C1-C2 = I totale - I intra-classe = I inter-classes

Le problème précédent revient donc, dans R^K muni de la métrique M (qu'il reste à discuter), à réaliser l'ACP du nuage N_I^*, ou, ce qui revient au même, de N_I. La représentation simultanée est obtenue en projetant les nuages N_I^* et N_I^j ; $j=1,J$ sur les axes obtenus.

On aboutit ainsi à une méthode dont le principe correspond à une pratique courante non justifiée théoriquement : réaliser l'ACP du tableau complet. Outre une formalisation plus précise des objectifs et une justification théorique, la méthode que nous proposons s'écarte (ou complète?) cette pratique sur plusieurs points dont voici les principaux :

- choix de la métrique M (i.e. pondération des groupes de variables)
- représentation simultanée ;
- représentation associée de l'interstructure ;
- aides à l'interprétation.

3.1- 5 Remarque sur la projection des nuages N_I^j

Le nuage N_I^j appartient au sous-espace $R^K{}_j$. Il peut paraître curieux de vouloir le projeter sur un vecteur de R^K qui n'appartient pas à $R^K{}_j$. Nous apportons ici quelques précisions sur ce point.

Notons :

u le vecteur de R^K issu de l'ACP de N_I ;

u_j la composante de u dans $R^K{}_j$ $\quad u = \sum_j u_j$;

F_j^I le facteur sur I qui correspond à la projection de N_I^j sur u ;

$\tilde{F}_j^I$ le facteur sur I qui correspond à la projection de N_I^j sur u_j ;

θ_j l'angle compris entre les vecteurs u et u_j.

.../...

On obtient aisément les relations suivantes :

$$\| u_j \| = \cos \theta_j$$

$$F_j^I = \cos \theta_j \, \tilde{F}_j^I$$

Les facteurs F_j^I et $\tilde{F}_j^I$ sont égaux à un coefficient près. Ainsi, lorsque l'on projette un nuage en dehors du sous-espace auquel il appartient, cela revient à réaliser successivement les opérations suivantes :

- projection sur un vecteur compris dans le sous-espace ;
- homothétie.

Le coefficient de l'homothétie, s'interprétant comme un cosinus, est toujours inférieur à 1 et introduit une contraction du nuage projeté. Ceci peut conduire à se demander s'il ne vaut pas mieux conserver les $\tilde{F}_j^I$ pour la représentation simultanée. En fait, il n'en est rien. Dans R^K, les axes u sont orthogonaux, ce qui n'est pas le cas des u_j. La prise en compte des $\tilde{F}_j^I$ conduit à superposer des nuages évoluant dans des espaces munis de métriques différentes, ce qui est illisible. A la rigueur, on pourrait le faire en se limitant à un seul axe u. Mais, même dans ce cas simple, la propriété qui veut que le nuage moyen soit au centre de gravité ne serait plus vérifiée. De plus, les points homologues ne resteront plus proches entre eux.

3.1-6 Un exemple "confetti"

Soit 3 individus A, B et C sur lesquels on a mesuré deux groupes contenant chacun une variable : V1 et V2. Le tableau de données est le suivant :

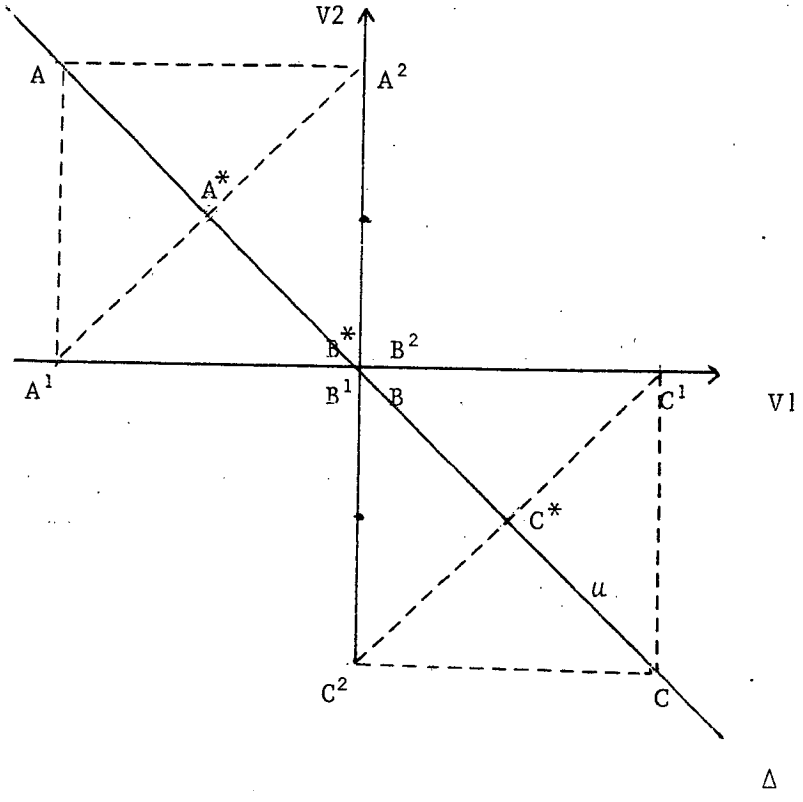
	V1	V2
A	-1	1
B	0	0
C	1	-1

Les structures sur les individus induites par les deux variables sont bien identiques (B est le centre de gravité de A et de C) quoique que cela n'apparait pas à première vue en raison d'une symétrie.

Dans R^K (ici $K=2$), nous noterons :

- A, B, C les points de N_I ;
- A^*, B^*, C^* les points de N_I^*
- A^1, B^1, C^1 les points de N_I^1 (A,B,C vus par V1)
- A^2, B^2, C^2 les points de N_I^2.

D'où la représentation dans R^K.



L'ACP de {A,B,C} conduit immédiatement à l'axe Δ qui est bien celui qui met le mieux en évidence l'identité des deux structures induites.

3.1- 7 Deux autres ACP dans R^K

Le fait de projeter l'ensemble des nuages N_I^j $j=1,J$ sur des vecteurs de R^K n'assure pas nécessairement une représentation simultanée intéressante. En effet, il est possible d'en imaginer d'autres a priori raisonnables.

.../...

ACP . de $\bigcup_j N_I^j$

Cette ACP fournit comme axes principaux les axes des analyses séparées des N_I^j. Ainsi, sur chaque axe, un seul des nuages N_I^j est représenté par des points différents de O.

Ce cas limite met en évidence le fait suivant : si on peut imaginer beaucoup de compromis entre la maximisation de C1 et la minimisation de C2, certains sont a priori inintéressants. Ici, on accepte de perdre sur C2 (proximités des points homologues), mais ce que l'on gagne sur C1 est, axe par axe, réparti très irrégulièrement sur les N_I^j.

Superposition des ACP partielles

On peut ainsi penser à réaliser les ACP partielles des J nuages N_I^j et prendre, comme directions de représentation simultanée les sommes des vecteurs principaux des ACP partielles. En réalité, ce procédé revient à superposer les graphiques des ACP partielles. On a cherché ici à maximiser C1 sans tenir compte de C2.

3.1-8 Résumé et Formules

Notations

N_I	nuage associé aux K variables
N_I^j	nuage associé aux K_j variables. Projection de N_I sur $\mathbb{R}^{k_j}$
N_I^*	$(1/J)N_I$ nuage moyen au centre de gravité des N_I^j
X	coordonnées de N_I
$\tilde{X}_j$	coordonnées de N_I^j dans $\mathbb{R}^k$ (X_j complété par des zéros)
u	axe d'inertie de N_I avec $u'Mu = 1$
F^I	$= XMu$ Projection de N_I sur u
F^{*I}	$= (1/J)\ F^I$ Projection de N_I^* sur u
F_j^I	$= \tilde{X}_j M u$ Projection de N_I^j sur u
$\cos \theta_j$	= cosinus de l'angle entre u et $\mathbb{R}^{k_j}$

Formules

(1) $W\ D\ F^I = \lambda F^I$

(2) $(1/J) \sum_j F_j^I = F^{*I}$

(3) $F_j^I = (1/\lambda)\ W_j\ D\ F^I$

(4) $\cos^2 \theta_j$ = contribution des variables du groupe K_j à la composante F^I

La formule (1) exprime simplement que F^I est une composante principale du tableau X d'inertie λ. La seconde indique que le nuage N_I^* est projeté au centre de gravité des N_I^j. La formule(3) est une relation entre F^I et F_j^I que nous allons démontrer.

Le vecteur u s'écrit en fonction de F^I :

$$u = (1/\lambda)\ X'\ D\ F^I$$

d'où :

$$\begin{aligned} F_j^I &= \tilde{X}_j\ M\ u \\ &= (1/\lambda)\ \tilde{X}_j\ M\ X'\ D\ F^I \\ &= (1/\lambda)\ W_j\ D\ F^I \end{aligned}$$

car il est facile de vérifier que $X_j\ M_j\ X'_j = \tilde{X}_j\ M\ X'$.

Démontrons la formule (4) qui montre que la contribution des variables du groupe K_j à F^I s'interprète dans $\mathbb{R}^k$. Cette contribution est, par définition, égale à la somme des contributions des variables du groupe :

$$CTR\ (K_j) = \sum_{k \in K_j} m_k\ (\text{projection de } v_k)^2\ /\lambda$$

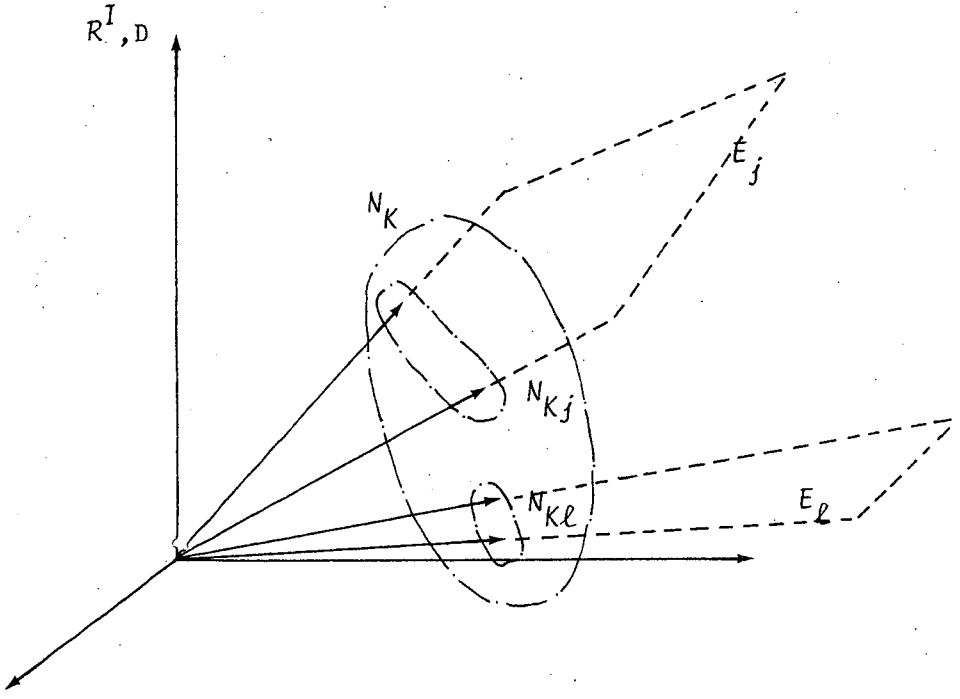
Or, la projection de la variable v_k sur l'axe associé à F^I dans l'ACP du tableau X, est le produit par $\sqrt{\lambda}$ de la k-ième coordonnée u_k de u . D'où :

$$\begin{aligned} CTR\ (K_j) &= \sum_{k\ \in K_j} m_k\ (u_k)^2 \\ &= \| \text{projection de } u \text{ sur } \mathbb{R}^{kj} \|^2 \\ &= \cos^2 \theta_j \end{aligned}$$

Il peut paraitre curieux de constater l'égalité entre le cosinus carré d'un angle et une contribution à l'inertie. Mais, en réalité, $\cos^2 \theta_j$ correspond aussi à la décomposition de l'inertie du nuage N_I projeté sur u, suivant les sous espaces orthogonaux $\mathbb{R}^{k_j}$ de $\mathbb{R}^k$.

3.2 - Interprétation dans R^I

3.2.1 - Rappel des notations (et quelques nouvelles)



- R^I est muni de la métrique D
- v_k vecteur représentant une variable
- N_K ensemble des vecteurs v_k
- N_{Kj} ensemble des vecteurs v_k appartenant au groupe j ($N_K = U_j N_{Kj}$)
- E_j sous espace engendré par les variables du groupe j

Dans cet espace nous étudions successivement les trois objectifs :

- représentation simultanée des nuages N_I^j
- étude des liaisons entre les groupes de variables
- comparaison des nuages de variables et de la direction de leurs axes d'inertie.

3.2.2 - Représentation simultanée des nuages N_I^j

Dans cette partie, nous présenterons la méthode que nous proposons de façon relativement indépendante de la partie précédente. Nous reprenons ainsi le problème presque à son point de départ en formulant nos objectifs dans R^I et en examinant quelle solution ce cadre de référence nous suggère pour les atteindre. Naturellement, cette solution conduit aux mêmes calculs que celle qui a été décrite dans R^K.

Cette démarche nous a paru plus intéressante qu'une simple traduction dans R^I des résultats issus des raisonnements dans R^K. En effet, le cadre R^K est, ainsi que nous l'avons déjà signalé, artificiel, et il était permis de se demander si la méthode proposée n'était pas strictement liée à ce cadre de référence. La présentation autonome que nous réalisons dans R^I montre qu'il n'en est rien.

Les propriétés demandées pour une bonne représentation sont toujours les mêmes mais se traduisent dans $\mathbb{R}^I$ de manière un peu différente. Nous cherchons toujours à superposer des projections des nuages N_I^j sur des espaces de petite dimension.

Projection sur un axe

Nous avons vu au § 2.4.4 qu'une projection de nuage N_I^j sur un axe peut être vue comme un vecteur V_j de R^I appartenant au sous espace E_j et à l'ellipsoïde $V_j' W_j^{-1} V_j = 1$

Plus précisément :

$$V_j' W_j^{-1} V_j = 1$$

ou

$$V_j' W_j^{-1} V_j \leq 1$$

suivant qu'il s'agit d'une projection sur un axe contenu dans le sous espace engendré par le nuage N_I^j ou extérieur à ce sous espace.

La recherche d'une représentation simultanée des nuages sur un axe peut donc être vue dans $\mathbb{R}^I$ comme la recherche de J vecteurs V_j appartenant aux ellipsoïdes $V_j' W_j^{-1} V_j$

P1 - Chacun des nuages est bien représenté

Pour une projection sur un axe, la propriété (P1) se traduit facilement.

En effet le carré de la norme de V_j est égal à l'inertie de la projection de N_I^j sur cet axe. Une projection de N_I^j donnant une bonne représentation de ce nuage est donc un vecteur de l'ellipsoïde $V_j' W_j^{-1} V_j = 1$ dont la norme est grande.

Les vecteurs de l'ellipsoïde dont la norme est maximum sont les vecteurs propres de $W_j D$ associés à sa plus grande valeur propre, i.e. la première composante principale du groupe de variables X_j . (cf § 2.4.4)

P2 - Les projections se ressemblent

Restant toujours dans le cas de projection sur un axe (nous étudions le cas général ensuite) cette propriété se traduit dans $\mathbb{R}^I$ par :

. les vecteurs V_j sont proches entre eux,

ou bien encore

. sont très correlés puisque la notion de proximité dans $\mathbb{R}^I$ correspond à celle de corrélation.

Compromis entre P1 et P2

Il est, bien évident qu'il faudra trouver un compromis entre ces deux propriétés, les premières composantes principales n'étant pas forcément très corrélées.

En imposant toujours aux vecteurs V_j d'appartenir aux ellipsoïdes, un compromis naturel entre :

. les normes sont maximum

. le cosinus de leur angle est maximum

est de chercher à obtenir des produits scalaires grands. En effet ces derniers étant le produit du cosinus par les normes de vecteurs, pour qu'il soit grand, il faut que tous les termes soient grands. On pourrait donc chercher à maximiser la somme de ces produits scalaires :

$$\Sigma_{j,j'} < V_j, V_{j'} >$$

Mais ce problème ne parait pas avoir de solution simple. D'où l'idée analogue à celle de CARROLL dans l'Analyse Canonique Généralisée de chercher à calculer les V_j en deux temps en utilisant un nuage moyen .

On construira donc un nuage moyen ; puis on cherchera une projection v de ce nuage sur un axe de grande inertie ; puis on cherchera des vecteurs V_j appartenant aux ellipsoïdes, qui soient à la fois proches de V et de norme grande. Pour ce dernier point on cherchera, pour les raisons évoquées ci-dessus, à rendre maximum le produit scalaire :

$$< V, V_j >$$

La projection V du nuage moyen sera proche des projections d'inertie importante de tous les nuages si les influences des différents nuages sont équilibrées. La proximité des V_j avec V assurera une certaine proximité des V_j entre eux. Le choix de V qui privilégie des directions de grande inertie et la maximisation du produit scalaire $<V, V_j>$ permettra d'obtenir des V_j de norme importante.

En procédant ainsi, la propriété suivante sera automatiquement vérifiée.

P3 - La représentation inclut un nuage moyen

Il faut donc construire un nuage représentant une "moyenne" des nuages d'individus associés à chaque groupe de variables. Il paraît naturel encore d'introduire le nuage associé à une moyenne pondérée de tous les groupes de variables. C'est le nuage associé à l'ensemble de toutes les variables, les variables de chaque groupe étant surpondérées par un coefficient α_j permettant d'équilibrer le rôle des différents groupes. Nous retrouvons le nuage moyen construit dans l'espace R^K.

La projection du nuage moyen d'inertie maximum est la première composante principale de ce nuage que l'on note V, c'est un vecteur propre de $\sum_j X_j W_j D$ associé à sa plus grande valeur propre. Sous cette forme, on peut voir que V est une compromis entre les directions de grande inertie de tous les nuages.

Pour ce nuage moyen, nous retrouvons donc les résultats obtenus dans $\mathbb{R}^K$.

Cherchons maintenant, à obtenir le vecteur V_j de l'ellipsoïde $V_j' W_j^{-1} V_j = 1$ rendant maximum le produit scalaire $\langle V_j , V \rangle$.

La solution est simple, V_j est à un coefficient près l'image de V par $W_j D$

En effet :

$$\begin{aligned} \langle V_j , V \rangle &= V_j' D V \\ &= V_j' W_j^{-1} W_j D V \end{aligned}$$

Cette expression peut être considérée comme le W_j^{-1} produit scalaire de V_j et de $W_j DV$ contrainte sur V_j peut se traduire avec la norme associée à ce produit scalaire : $||V_j|| = 1$. Le vecteur V_j rendant maximum le produit scalaire est donc proportionnel à $W_j DV$:

$$V_j = \alpha \; W_j DV$$

Pour calculer le coefficient α, il suffit d'écrire :

$$\begin{aligned} 1 &= V_j' W_j^{-1} V_j \\ &= \alpha^2 V' D (W_j W_j^{-1}) W_j) DV \\ &= \alpha^2 \langle V, W_j DV \rangle \end{aligned}$$

d' où $$V_j = \frac{W_j D V}{\sqrt{\langle v, W_j DV \rangle}}$$

Nous avons donc obtenu des vecteurs V_j qui sont des projections des nuages N_I^j d'assez grande inertie et liées entre elles. Mais, utiliser directement ces vecteurs V_j pour une représentation simultanée des nuages pose plusieurs problèmes :

1. la norme des V_j est indépendante de la liaison entre V et V_j, et ceci risque de donner une présentation peu claire, les V_j très peu liés au vecteur moyen V risquant de brouiller les résultats. Il serait préférable de diminuer la norme de ces derniers.

2. la projection d'un point du nuage moyen ne va pas se trouver au centre de gravité de ses homologues dans les N_I^j : $1/J \sum_j V_j \neq V/J$, propriété facilitant beau-

coup la comparaison

3. le troisième problème est d'ordre tout à fait différent. Jusqu'ici nous n'avons étudié que des projections sur un axe, mais il faut construire des projections sur des plans ou des espaces de dimension supérieure. Pour cela, on peut itérer le procédé, chercher une projection de grande inertie du nuage moyen sur un axe orthogonal au premier. On obtiendra bien sûr la seconde composante principale de ce nuage. Les projections des N_I^j liées à cette seconde composante peuvent s'obtenir comme pour la première. Mais il n'y a aucune raison que pour chaque nuage ces projections représentent des projections sur des axes orthogonaux aux premiers. Pour représenter simultanément les nuages sur une même base orthogonale, il faudrait introduire des métriques différentes pour chaque nuage ce qui est illisible et perd totalement l'intérêt d'une représentation simultanée.

La solution est de prendre non pas V_j mais :

$$\tilde{V}_j = \frac{1}{\lambda} W \quad DV$$

où λ est l'inertie de la projection V du nuage N_I

On résoud ainsi tous les problèmes :

. pour le second il est facile de voir que :

$$\sum_j \tilde{V}_j = (1/\lambda) \sum_j W_j \quad DV$$

$$= (1/\lambda) \quad WDV$$

$$= V$$

. le premierest résolu aussi puisque, pour obtenir $\tilde{V}_j$, nous avons multiplié V_j par le produit scalaire $\langle V, V_j \rangle$ qui représente la liaison entre V et V_j :

$$\tilde{V}_j = (1/\lambda) \quad \sqrt{\langle V, W_j \, DV \rangle} \; V_j$$

et

$$\langle V, V_j \rangle = \langle V, W_j \quad DV \rangle \qquad / \qquad \sqrt{\langle V, W_j \, DV \rangle}$$

$$= \sqrt{V, W_j \, DV}$$

. Quand au troisième problème, pour voir qu'il est résolu, il suffit de remarquer qu'on obtient avec ces $\tilde{v}_j$ exactement la même solution que dans $\mathbb{R}^K$. Nous pouvons donc considérer deux vecteurs $\tilde{v}_j$ et $\tilde{v}_j'$ déduits de deux composantes différentes comme des projections sur des axes orthogonaux qui n'appartiennent pas au sous espace engendré par les nuages N_I^j.

La solution proposée dans $\mathbb{R}^K$ se justifie donc dans $\mathbb{R}^I$, nous verrons qu'elle se traduit facilement aussi dans $(\mathbb{R}^I)^*$

3.2.3 - Etude des liaisons entre groupes de variables

Rappelons qu'au paragraphe 1.6 nous avions suggéré de chercher une généralisation de l'analyse des correspondances multiples à des groupes de variables quelconques qui :

a) permette de calculer des variables canoniques qui expriment une part assez importante de l'inertie des groupes de variables (ceci n'est pas toujours le cas pour l'analyse multicanonique au sens de CARROLL qui se confond avec l'analyse des correspondances multiples dans le cas de variables qualitatives).

b) permette une représentation des deux ensembles, individus et variables, du type analyse en composantes principales.

Le principe de l'analyse multicanonique au sens de Carrol est de chercher d'abord une variable générale la plus liée possible à tous les groupes plus précisément une variable u rendant maximum la somme des carrés des coefficients de corrélation multiple entre u et chaque groupe de variables. Géométriquement, dans R^I, ceci revient à chercher un vecteur rendant maximum la somme des cosinus carrés des angles entre u et les sous-espaces E_j engendrés par chaque groupe de variables. Cette variable u étant obtenue, on cherche pour chaque groupe, la combinaison linéaire des variables du groupe la plus corrélée à u. C'est la projection orthogonale de u sur le sous-espace E_j.

On obtient donc une variable u et un J-uplets de variables u_j corrélées entre elles. Notons θ_j d'angle entre u et u_j, la quantité que l'on a maximisé en $\sum_j \cos\theta^2{}_j$

On cherche ensuite une seconde variable u_2 orthogonale à u et rendant maximum le même critère, puis on projette u_2 sur les sous-espaces E_j et on itère le procédé.

On montre [cf 8et14]que la solution est donnée par la diagonalisation de la somme des opérateurs de projection orthogonale sur les sous-espaces E_j . Les variables générales sont les vecteurs propres de cette somme, ordonnées dans l'ordre décroissant des valeurs propres (qui sont égales aux quantités maximisées $\sum_j \cos^2\theta_j$).

Il est évident que dans cette technique, seuls entrent en ligne de compte les sous-espaces E_j engendrés par les groupes de variables. Ce qui fait que l'on peut obtenir comme premières variables canoniques des combinaisons linéaires des variables d'un groupe exprimant une variance très faible de ce groupe.

Pour éviter cet inconvénient il est nécessaire d'introduire la notion de variance du groupe exprimée par un vecteur du sous espace E_j , ou ce qui est équivalent d'inertie de la projection du nuage de variables X_j dans une direction de E_j .

Or dans le § 2.5 nous avons introduit :

. *le tenseur* $W_j = X_j' M_j X_j$ qui contenait cette information

. *une mesure de liaison* entre deux groupes de variables définie à l'aide de ce tenseur. Cette mesure de liaison qui tient compte de l'inertie du groupe de variable dans les différentes directions s'écrivait si le second groupe était réduit à une seule variable u :

$$< W_j , u \otimes u > = \Sigma_{k \in j} \; m_k < u , v_k >^2$$

= Inertie de la projection des v_k *sur* u

= Trace (W_j Du u' D)

Dans le cas de variables qualitatives (cf § 2.5.3.d) cette mesure de liaison se confond avec le carré du coefficient de corrélation multiple $cos^2 \theta_j$.

Si nous procédons d'une manière tout à fait analogue à celle proposée par Carroll, en remplaçant $cos^2 \theta_j$ par la mesure de liaison ci-dessus, nous obtiendrons, si le problème se résoud, une variante de l'analyse multicanonique de Carroll, tenant compte de l'inertie des groupes dans les différentes directions et se confondant avec l'analyse des correspondances multiples dans le cas qualitatif.

Montrons maintenant que le problème se résoud très facilement avec cette nouvelle mesure de liaison. Suivant le principe de Carroll, on cherche d'abord une variable générale u rendant maximum la somme des liaisons de u avec chaque groupe. Nous cherchons donc u rendant maximum la quantité :

$$T(u) = \Sigma_{j \in J} \; \Sigma_{k \in K_j} \; m_k < u, v_k >^2$$

= Inertie des projections des v_k *sur* u

Variables générales

La solution est simple, le vecteur u qui rend maximum l'inertie des projections des v_k sur u est la première composante principale de ces variables, i.e. un vecteur propre de WD associé à sa plus grande valeur propre. Les vecteurs propres de WD étant D-orthogonaux, le vecteur u_2 orthogonal à u maximisant l'inertie des projections des v_k est la seconde composante principale des v_k, i.e. le second vecteur propre de WD etc...

Nous obtiendrons donc comme variables générales de l'analyse des liaisons les composantes principales du tableau X dans l'ordre décroissant de leur inertie. Le calcul est donc très simple, et surtout nous obtiendrons la représentation classique des variables et des individus de l'analyse en composantes principales. Les résultats sont donc très facilement interprétables.

Remarquons que jusqu'ici, l'analyse des liaisons que nous proposons peut se déduire de celle de Carroll en remplaçant l'opérateur de projection P_j sur le sous espace E_j engendré par le groupe de variables j par l'opérateur W_jD qui a même image mais qui tient compte de l'inertie des variables dans les différentes directions de E_j(cf. § 2.5.1.c).

En effet :

$$\langle u, P_j u \rangle_D = \cos^2 \theta_j$$

$\langle u, W_jDu \rangle_D$ = *liaison entre u et le groupe de variables j que nous utilisons*

Dans l'analyse multicanonique, on cherche un vecteur u normé rendant maximum

$$\sum_j \cos^2 \theta_j = \sum_j \langle u, P_j u \rangle_D$$

$$= \langle u, \sum_j P_j u \rangle_D$$

Nous cherchons u normé rendant maximum :

$$\sum_j \langle u, W_j D u \rangle_D = \langle u, \sum_j W_j D u \rangle_D$$

Dans le premier cas, la solution est donnée par la diagonalisation de $\sum_j P_j$ et dans le second par celle de $\sum_j W_j D$.

Il est bien clair que dans le second cas les variables générales sont beaucoup plus attirées par les directions de grande inertie des nuages de variables.

Variables canoniques

Les variables générales obtenues, il reste à calculer les variables canoniques, i.e. les combinaisons linéaires des variables de chaque groupe les plus liées aux variables générales obtenues. Dans l'analyse multicanonique au sens de Carroll, ce sont tout simplement les projections de ces variables sur les E_j, i.e. les variables des E_j les plus corrélées à u.

Nous pourrions choisir la même définition mais elle serait peu cohérente avec la mesure de liaison que nous avons choisie et ne répondrait pas à notre souci d'obtenir des variables exprimant une part assez importante de la variance du groupe j.

Or, nous venons de voir que dans le calcul des variables générales, nous avons remplacé les projecteurs P_j par les opérateurs $W_j D$. Il est donc logique, pour favoriser les directions de grande inertie, de remplacer aussi ces projecteurs par $W_j D$ dans le calcul des variables canoniques. Par définition, la variable canonique du groupe j associée à la variable générale sera donc $W_j D u$.

Le choix de $W_j D u$ se justifie car il correspond bien aux buts que nous nous étions fixés dans l'étude des liaisons entre les groupes de variables. En effet, les variables obtenues extraient dans chaque groupe une part de variance assez importante et la méthode est une généralisation de l'analyse multicanonique qui se confond avec l'analyse des correspondances multiples dans

le cas des variables qualitatives (puisque pour une variable qualitative $W_j D$ est égal à P_j (cf. § 2.5.1.d). Mais surtout, nous retrouvons avec ces $W_j Du$ les projections des nuages N_J^j proposées pour leur représentation simultanée au § précédent, ce qui permet d'interpréter ces variables.

Cas où tous les groupes sont identiques

Il est souvent intéressant pour juger de l'intérêt d'une méthode de l'appliquer à un cas limite même si ce dernier risque fort peu de se produire en réalité. Ceci permet de voir si le résultat est logique pour ce cas limite et pour les cas réels s'en approchant.

Si tous les groupes sont identiques, les variables générales sont les composantes principales de ce groupe, qui donnent une description plus complète de ce groupe que les vecteurs d'une base quelconque orthogonale du sous espace engendré. Les variables canoniques sont confondues avec les variables générales. En effet, elles sont obtenues en appliquant $W_j D$ aux variables générales. Ces dernières sont vecteurs propres de WD et donc des $W_j D$ qui sont tous égaux.

En résumé :

u = vecteur rendant maximum la somme de ses liaisons avec tous les groupes
= vecteur propre de $\Sigma_j W_j D$
= première composante principale de toutes les variables
= projection du "nuage moyen" N_I^*

$u_j = W_j D u$
= variable canonique associée à la variable générale
= projection de N_I^j permettant la représentation simultanée de ces nuages

3.2.4 - Comparaison des nuages de variables

Nous avions suggéré au § 1.5 de comparer les directions des axes d'inertie des différents nuages de variables. En effet, ces axes d'inertie sont des éléments caractéristiques de chaque tableau, puisqu'ils se confondent avec les composantes principales. Les composantes étant centrées, leur coefficient de corrélation est égal au cosinus de leur angle dans $\mathbb{R}^I$.

Pour comparer rapidement toutes ces composantes, nous proposons de projeter leur vecteur unitaire sur un espace de petite dimension bien ajustée. Ainsi on repérera facilement les composantes semblables, celles qui sont caractéristiques d'un ou plusieurs groupes, etc...

le poids de chaque composante sera donc sa valeur propre.

Il reste à déterminer l'espace sur lequel nous les projetterons. Dans le contexte d'analyse factorielle qui est le nôtre, il est naturel de chercher un ajustement aux moindres carrés de l'ensemble de ces composantes. Il est naturel aussi d'accorder plus d'importance aux premières composantes qui contiennent davantage d'information et qui nous intéressent donc plus. Pour cela, il suffit d'affecter dans l'ajustement aux moindres carrés un poids à chaque composante, croissant avec l'importance de cette composante. Or, cette importance est mesurée par la valeur propre associée, i.e. l'inertie de la projection du nuage de variables sur cet axe, Le poids de chaque composante sera donc sa valeur propre.

La solution de ce problème est donnée par l'analyse en composantes principales de toutes ces composantes ainsi pondérées. Or, nous allons montrer que cette analyse est équivalente à celle du nuage de toutes les variables. Il suffira donc, dans cette analyse que nous avons déjà proposée pour atteindre les autres objectifs, de mettre en éléments supplémentaires les composantes principales de chaque groupe de variables.

Démontrons maintenant l'équivalence de l'analyse en composantes principales de toutes les variables et des composantes principales de tous les groupes (chaque composante normée ayant pour poids son inertie). Notons :

H_j *la matrice des composantes normées du groupe j*

Δ_j *la matrice diagonale des valeurs propres associées*

Les composantes sont les vecteurs propres de $W_j D$ qui s'écrit donc :

$$W_j D = H_j \Delta_j H_j' D$$

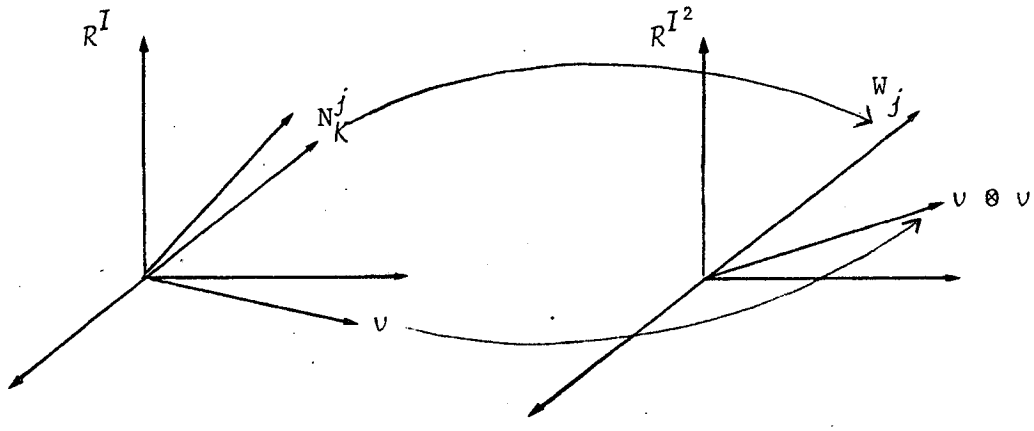
La matrice d'inertie du groupe j est donc égale à celle de ses composantes principales $H_j \Delta_j H_j'$.

La matrice d'inertie de toutes les variables, qui est la somme des matrices W_j est donc égale à la matrice d'inertie des composantes principales de tous les groupes. Les résultats des deux analyses sont donc identiques.

3.3 La méthode dans R^{I^2}

Dans la première partie, nous avons indiqué les objectifs sous-jacents à l'étude simultanée de plusieurs tableaux. Nous avions alors rassemblé plusieurs questions sous les termes désormais classiques d'étude de l'interstructure et d'étude des intrastructures. Les deux paragraphes précédents, qui exploitaient les cadres de référence R^K et R^I ont présenté une méthode d'étude des intrastructures. Nous utilisons maintenant le cadre R^{I^2} pour introduire une méthode d'étude de l'interstructure.

3.3-1 Rappel des notations



N_K^j nuage des vecteurs représentant les variables du groupe j

v vecteur de R^I

W_j tenseur d'inertie associé à N_K^j

$v \otimes v$ tenseur d'inertie associé à v

N_J ensemble des W_j

3.3-2 Le problème de la représentation de l'interstructure

Dans R^{I^2}, chaque groupe de variables est un vecteur. Dans cet espace,

.../...

nous avons défini (§ 2.5-2) une métrique euclidienne qui permet de calculer des distances entre les tenseurs W_j . Ainsi, nous avons montré que le produit scalaire entre W_j et $W_{j'}$ est une mesure de liaison entre les groupes de variables j et j'. Dans R^{I^2}, la forme du nuage N_J est l'interstructure.

Comme toujours en analyse factorielle, nous cherchons à décrire le nuage N_j en le projetant dans un espace de faible dimension. On obtient alors une représentation de l'interstructure qui, pour être utilisable, doit posséder les propriétés suivantes.

(pa) *Le nuage N_J doit être bien représenté*

On retrouve ici le problème général de l'analyse factorielle. Remarquons simplement que les angles entre tenseurs nous intéressent et qu'il ne convient pas de centrer le nuage de points.

Cette propriété est généralement la seule qui est retenue. Elle conduit à une ACP du nuage N_J, les éléments étant considérés comme des variables. L'inconvénient, important à nos yeux, de ce type d'analyse est de fournir un repère constitué d'axes difficilement interprétables. D'où la prise en considération des propriétés suivantes.

(pb) *Les axes du repère doivent être interprétables en termes de variables*

Cette propriété implique que les axes du repère soient des tenseurs symétriques de rang 1. Ils sont alors induits par un seul vecteur de R^I. Ce vecteur est lui-même interprétable à l'aide des angles qu'il constitue avec chacune des variables.

.../...

(pc) Les axes du repère doivent être liés à ceux de la représentation simultanée des intra-structures

La représentation de l'interstructure met en évidence des ressemblances (et dissemblances) entre groupes de variables. Ces ressemblances résument, à l'aide d'un indicateur unique, les différences entre les intra-structures. Ainsi, la représentation de l'interstructure ne peut être dissociée de la représentation simultanée des intrastructures. Plus précisément, ces deux représentations doivent mettre en évidence les mêmes phénomènes.

3.3-3 La méthode

Pour étudier le nuage N_J dans R^{I^2}, nous le projetterons sur un sous-espace de petite dimension. Comme toujours en analyse factorielle, nous construirons ce sous-espace progressivement en déterminant d'abord un vecteur de base, puis un second orthogonal au premier, et ainsi de suite. Nous détaillerons surtout la recherche du premier vecteur.

Les propriétés *(pa)* et *(pb)* posent le problème de la façon suivante : chercher le tenseur symétrique normé de rang 1 noté $v \otimes v$ ($v \in R^I$)" ajustant bien"le nuage.

Le critère d'ajustement que l'on choisit habituellement dans une telle situation est celui de la somme des carrés des projections. Or, la situation que nous analysons ici présente (entre autres) deux particularités importantes pour notre propos :

1/ $\langle v \otimes v, w \otimes w \rangle = cos^2\ \Theta(v,w)$ en appelant $\Theta(v,w)$ l'angle entre v et w dans R^I (cf § 2.5.3 a).

2/ $\langle W_j, v \otimes v \rangle \geq 0$; le produit scalaire entre W_j et un tenseur symétrique de rang 1 est positif ou nul (cf § 2.5.3-b).

.../...

Compte tenu de la cohérence souhaitée entre les démarches dans R^I et R^{I^2} (propriété (pc), la première particularité suggère d'utiliser les projections (et non leur carré) dans le critère d'ajustement. La deuxième particularité nous indique que le critère "somme des projections" a un sens.

Ces considérations nous conduisent à rechercher $v \otimes v$ tel que $\sum_j \alpha_j \langle W_j , v \otimes v \rangle$ maximum. Etant donné que $W_j = \sum_{k \in K_j} m_k v_k \otimes v_k$, le critère devient :

$$\sum_j \alpha_j \sum_{k \in K_j} m_k \langle v_k \otimes v_k, v \otimes v \rangle = \sum_{k \in K} \alpha_j m_k \langle v_k \otimes v_k, v \otimes v \rangle$$

ce qui s'interprète dans R^I comme $\sum_{k \in K} \alpha_j m_k \cos^2 \theta(v_k, v)$.

Nous retrouvons le critère C du § 3.2-2. Le vecteur v de R^I est la première composante principale associée au tableau entier; le tenseur cherché est $v \otimes v$, associé à cette première composante principale.

Ce résultat aurait pu être posé a priori : on aurait ainsi pu décider de projeter le nuage N_J sur les tenseurs associés aux vecteurs trouvés dans R^I et, indirectement dans R^K. Mais dans ce cas, la représentation de l'interstructure serait apparue seulement comme une "aide à l'interprétation" de la représentation simultanée des intra-structures. On se serait alors préoccupé uniquement de la propriété (pc). Une telle présentation aurait suggéré que, en posant le problème d'abord dans R^{I^2}, puis dans les autres cadres, on aurait abouti à d'autres résultats.

En réalité, la représentation de l'interstructure que nous proposons est optimale en elle-même. Naturellement, cette optimalité n'est pas unique, et d'autres méthodes adoptent d'autres critères. En particulier, des ACP di-

.../...

rectes du nuage N_J ont été proposées. Entre ces deux démarches, la différence essentielle, si on se limite au cadre R^{I^2}, est de choisir des tenseurs symétriques de rang 1 (et non de rang quelconque), pour définir le sous-espace de représentation de N_J. Ce choix résulte d'un compromis entre les propriétés *(pa)* et *(pb)* : par rapport à l'ACP, on perd en qualité de représentation, mais l'on gagne en facilité d'interprétation.

La recherche d'un deuxième vecteur, qui optimise le même critère, mais qui satisfait à la contrainte d'orthogonalité avec le premier, ne pose pas de problème particulier. Ce vecteur est le tenseur de rang 1 associé au deuxième vecteur issu de l'ACP dans R^I. Nous avons vu (§ 2.5.3.4.) que si dans R^I $<v_1, v_2> = 0$, alors, dans R^{I^2} $<v_1 \otimes v_1, v_2 \otimes v_2> = 0$.

D'un point de vue pratique, les calculs nécessités par la représentation de l'interstructure se déduisent directement des résultats de l'ACP dans R^I. En effet, si v est un vecteur de R^I, la projection de $W_j D$ sur $v \otimes v$ est égale à la somme des inerties des projections des variables du groupe j sur v. (cf § 2.5.3.b).

3.3-4 Remarques sur l'interprétation

La représentation de l'interstructure peut être vue en elle-même (elle est optimale en un certain sens) et en tant qu'aide à l'interprétation de la représentation simultanée des intra-structures. Nous distinguerons ici ces deux points de vue.

3.3-4 1 La proximité entre deux tenseurs

Dans R^{I^2}, la proximité de deux tenseurs traduit bien, nous l'avons vu au § 2.5.3 la ressemblance entre deux groupes de variables.

.../...

En projection, cette proximité se traduit de la même façon, à condition que les vecteurs projetés soient bien représentés. Or, la base de projection est très particulière puisqu'elle est formée uniquement de tenseurs symétriques de rang 1. Un tenseur de rang r étant une combinaison linéaire de r tenseurs de rang 1, il faudrait déjà r composantes - au moins - pour qu'il soit bien représenté.

De plus, certains tenseurs ne seront jamais bien représentés, même si l'on augmente le nombre de composantes. En effet, le sous-espace engendré par les tenseurs de la forme $v \otimes v$ (avec $v \in R^I$) est beaucoup plus petit que R^{I^2}, il est au plus de dimension n et ne contient que des tenseurs de la forme :

$$a_1 v_1 \otimes v_1 + \ldots \qquad a_s v_s \otimes v_s$$

Or, tous les tenseurs d'inertie ne s'écrivent pas sous cette forme. Prenons par exemple le tenseur d'inertie d'une variable normée u située dans le plan v_1, v_2.

$$u = x_1 v_1 + x_2 v_2$$

Le tenseur associé à u s'écrit :

$$u \otimes u = x_1^2 v_1 \otimes v_1 + x_2^2 v_2 \otimes v_2 + x_1 x_2 v_1 \otimes v_2 + x_1 x_2 v_1 \otimes v_2$$

Ce tenseur n'appartient généralement pas au plan $v_1 \otimes v_1, v_2 \otimes v_2$ ni même à un sous-espace engendré par un nombre quelconque de composantes principales car des termes "non diagonaux" ne sont pas nuls.

La qualité de représentation des tenseurs sera donc généralement assez mauvaise et il sera difficile de conclure à la proximité totale de deux tenseurs. Par contre, seront bien mises en évidence sur les graphiques les structures communes à tous les groupes de variables, ou spécifiques de certains d'entre eux.

.../...

3.3-4 2 La représentation de l'interstructure en tant qu'aide à l'interprétation

Nous avons vu, dans une remarque reliant les calculs nécessités par la représentation de l'interstructure (§....), que la coordonnée d'un point W_j sur un axe $v \otimes v$ de R^{I^2} était égale à la somme des inerties des variables du groupe j le long de la direction v associée dans R^I. A un coefficient près (i.e l'inertie totale dans cette direction), la coordonnée de W_j est égale à la somme des contributions absolues des variables du groupe j à v. Cette coordonnée est la contribution absolue du groupe j à v. Elle mesure l'importance de ce groupe dans la détermination de v. Cette propriété permet de considérer la représentation de l'interstructure comme une aide à l'interprétation des analyses dans R^I et R^K.

La coordonnée de W_j admet pour maximum sa plus grande valeur propre. On retrouve ici, sous un autre aspect que dans le § 3.5., l'intérêt de rendre cette plus grande valeur propre égale à 1. On équilibre ainsi le rôle des différents groupes dans la détermination des composantes principales.

3.3.5 - Le modèle INDSCAL

Rappelons (cf §1-7) que le modèle INDSCAL suppose que les distances entre les individus définies par les différents groupes peuvent se décomposer suivant des "facteurs" communs à tous les groupes, le poids affecté à chaque facteur différent suivant les groupes.

Notons

$F_s(i)$ la valeur du s-ième facteur pour l'individu i

w_s^j le poids affecté à F_s par le groupe j

S le nombre total de facteurs

Y la matrice de dimension I x S ayant en colonne les facteurs F_s

Δ_j la matrice S x S diagonale ayant comme éléments diagonaux les poids w_s^j affectés aux facteurs par le groupe j

W_j la matrice I x I des produits scalaires entre individus définis par le groupe j.

Le modèle s'écrit :

$$d_j^2(i,i') = \sum_s w_s^j \{F_s(i) - F_s(i')\}^2 + \mu_{ii'}^j$$

ou bien :

$$W_j = Y \Delta_j Y' + \varepsilon_j$$
$$= \sum_s w_s^j F_s F_s' + \varepsilon_j$$

Ce modèle s'applique dans un cas plus général que le nôtre où l'on ne dispose pas de matrices de coordonnées, mais seulement de matrices de proximité. Nous nous intéressons seulement à son application dans notre cas particulier.

Le modèle posé, il reste à calculer les paramètres qui l'ajustent le mieux aux données réelles. Les paramètres sont ici les facteur F_s et les poids w_s^j.

Les critères d'ajustement sont nombreux ils cherchent à minimiser l'écart entre les carrés des distances dans le modèle et dans les données réelles (stress),

ou bien à minimiser l'écart entre les produits scalaires réels et ceux du modèle. Dans ce deuxième cas , très fréquent, la quantité à minimiser s'écrit en notant $\tilde{W}_j$ les produits scalaires obtenus par le modèle :

$$V = \sum_j ||\varepsilon_j||^2 = \sum_j ||W_j - y_\Delta y'||^2$$

$$= \sum_j \text{Trace } (W_j - \tilde{W}_j)^2$$

où $//.//^2$ est la norme dans $(\mathbb{R}^I)^2$ lorsque la métrique D est l'identité. L'introduction des poids des individus dans la métrique D permet de donner une forme un peu plus générale au critère.

Les $\tilde{W}_j$ se déduisent du modèle et s'écrivent donc en fonction des facteurs F_Δ et des poids w_Δ^j (que l'on doit calculer).

$$\tilde{W}_j = \sum_\Delta w_\Delta^j \; F_\Delta \otimes F_\Delta$$

Dans $(\mathbb{R}^I)^2$, $\tilde{W}_j$ est donc une combinaison linéaire des tenseurs $F_\Delta \otimes F_\Delta$ et appartient au sous-espace qu'ils engendrent.

Projection des W_j et calcul des w_Δ^j

Supposons maintenant que les F_Δ soient déjà calculés et que l'on cherche les poids w_Δ^j minimisant ν. Pour chaque W_j on cherche donc les coefficients w_Δ^j qui rendent minimum.

$$// \; W_j - \sum_\Delta w_\Delta^j \; F_\Delta \otimes F_\Delta \; //$$

Il est évident géométriquement que la solution est donnée par la projection de W_j sur le sous-espace engendré par les $F_\Delta \otimes F_\Delta$.

Si les facteurs F_Δ sont orthonormés, les tenseurs $F_\Delta \otimes F_\Delta$ le seront aussi, et la projection de W_j sur le sous espace qu'ils engendrent se calcule très simplement :

w_Δ^j est la coordonnée de la projection de W_j sur $F_\Delta \otimes F_\Delta$.

donc, w_Δ^j = inertie des projections du groupe de variable j sur F_Δ dans $\mathbb{R}^I$ (cf §2-5-2).

(L'orthogonalité des F_Δ est une contrainte souvent imposée dans les solutions proposées, et la normalisation ne fait que lever une indétermination).

Les poids w_δ^j peuvent donc s'interpréter géométriquement dans $(\mathbb{R}^I)^2$ et dans $\mathbb{R}^I$.

Si, de plus, les F_δ sont les composantes principales de l'ensemble de toutes les variables, $\tilde{W}_j$ est la projection de W_j sur le sous espace engendré par les tenseurs de rang un associés à chaque composante. Le tenseur $\tilde{W}_j$ obtenu sera alors exactement celui que nous avons introduit dans le paragraphe précédent pour comparer les tenseurs entre eux. Les projections peuvent donc s'interpréter comme les tenseurs des représentations approchées des nuages N_I^j suivant le modèle INDSCAL. Les facteurs communs des modèles sont les composantes principales F_δ^I du nuage moyen et les coordonnées des W_j sur $F_\delta^I \otimes F_\delta^I$ sont les poids w_j^δ affectés au facteur F_δ par le groupe j.

On retrouve le fait que la plupart des tenseurs ne sont jamais bien représentés par leur projection sur les sous espaces engendrés par les $F_\delta \otimes F_\delta$ puisque les nuages correspondant à ces tenseurs projetés sont d'un type très particulier : leurs projections sur le s-ième axe de la base sont égales à $\sqrt{w_\delta^j}\, F_\delta$, ce ne sont pas des projections des nuages réels N_I^j que l'on pourrait approcher et atteindre en augmentant le nombre de composantes.

Interprétation dans $\mathbb{R}^K$

Nous avons donc proposé deux représentations simultanées différentes des nuages N_I^j basées sur la même représentation du nuage moyen sur ses axes principaux d'inertie. Comparons-les.

La première est une projection du nuage N_I^j. Rappelons que sa projection sur l'axe δ de la base commune est :

$$F_j^I = (1/\lambda_\delta)\, W_j\, D F_\delta^I$$

Rappelons aussi que F_j^I est une projection sur un axe qui n'est pas contenu dans le sous-espace engendré par N_I^j. Dans $\mathbb{R}^K$, c'est la projection sur un axe u; cet axe u se projette sur $\mathbb{R}^{K_j}$ suivant un axe u_j qui est contenu dans le sous-espace engendré par N_I^j. En notant, comme au §3-1, $\cos\theta_j$ l'angle entre u et u_j et $\tilde{F}_j^I$ la projection de N_I^j sur u_j, on a :

$$(1) \qquad F_j^I = \cos\theta_j\, \tilde{F}_j^I$$

La seconde représentation simultanée des N_I^j est basée sur le modèle INDSCAL. Ce n'est pas une projection orthogonale des N_I^j. Notons $\bar{F}_j^I$ la projection du nuage représentant N_I^j sur l'axe δ de la base commune, (identique à la précédente). Elle est homothétique de F_δ^I, de rapport d'homothétie $\sqrt{w_\delta^j}$ Or w_δ^j est la coordonnée, de la projection de W_j sur $F_1^I \otimes F_1^J$ dans $(R^I)^2$. Nous avons montré que cette coordonnée était la somme des inerties des projections des variables du groupe j sur F_δ^I et au §3-1-7 nous avons montré que ceci était égal à $\cos^2\theta_j$. D'où

$$(2) \qquad \bar{F}_j^I = \cos\theta_j\, F_\delta^I$$

Comparons (1) qui est la projection de N^j dans la première représentation simultanée et (2) qui est la projection sur le ^{I}même vecteur de base du nuage représentant N_I^j suivant le modèle INDSCAL. On peut interpréter le résultat donné par INDSCAL comme une approximation d'une projection de N_I^j consistant à remplacer la projection $\tilde{F}_j^I$ de N_I^j sur un axe du sous-espace qu'il engendre par la projection F_δ^I du nuage moyen (sachant que les projections F_δ^I et $\tilde{F}_j^I$ sont très liées puisque F_δ^I est la projection du nuage moyen sur un axe u de $\mathbb{R}^k$ et $\tilde{F}_j^I$ la projection de N_j^I sur la projection sur $\mathbb{R}^{K_j}$ de u).

Cette interprétation est intéressante car elle permet de comparer, axe par axe, pour chaque nuage N_I^j la représentation donnée par INDSCAL et le nuage réel. Ce qui est une comparaison beaucoup plus fine qu'un simple indice global mesurant l'écart entre le tenseur du nuage réel et le tenseur du nuage représenté par INDSCAL.

Le choix des facteurs F_δ^I

Nous pouvons donc proposer une solution pour le modèle INDSCAL. Dans cette solution, les facteurs F_δ^I étant imposés, les w_δ^j minimisent ε_j le carré de la norme de la différence entre la matrice des produits scalaires réels et celle du modèle ε_j Mais, les facteurs F_δ^I sont-ils bien choisis , minimisent-ils$V$? Nous l'étudions ici.

Les $\tilde{W}_j$ étant les projections des W_j sur le sous-espace engendré par les $F_\delta^I \otimes F_\delta^I$, v peut s'**inscrire**

$$v = \sum_j \| W_j - \tilde{W}_j \|^2$$

$$= \sum_j \| W_j - \text{Projection de } W_j \text{ sur } \{F_1^I \otimes F_1^I, \ldots, F_\delta \otimes F_\delta^I\} \|^2$$

Minimiser v, revient à trouver un sous espace de $(\mathbb{R}^I)^2$ engendré par des tenseurs symétriques de rang un ajustant au mieux les W_j au sens des moindres carrés.

Si on impose aux facteurs d'être orthogonaux deux à deux le problème peut se poser de manière équivalente en cherchant successivement les F_1^I, $F_2^I, \ldots F_\delta^I$.

On cherche d'abord F_1^I normé tel que soit maximum :

$$v_1 = \sum_j \| \text{Proj. de } W_j \text{ sur } F_1^I \otimes F_1^I \|^2$$

$$= \sum_j (\text{inertie des projections des variables du groupe } j \text{ sur } F_1^I)^2$$

Puis on cherchera F_2^I orthogonal à F_1^I maximisant le même critère etc...

Le F_1^I que nous obtenons dans l'analyse en composantes principales du tableau entier ne maximise pas ce critère, il maximise la somme des inerties des variables et non la somme des carrés des inerties. Il est donc certain que la solution que nous proposons ne minimise pas le critère ν.

Cette solution présente cependant un certain intérêt. Tout d'abord elle est extrêmement simple et les calculs sont très courts. D'autre part, et surtout, tous les autres résultats donnés par la méthode permettent d'enrichir l'interprétation.

Pour voir si cette méthode est raisonnable, il faut cependant que l'on ne s'éloigne pas trop du critère à optimiser avec les composantes principales du nuage moyen. Plaçons nous d'abord dans le cas où le modèle est exact, nous allons voir que nous obtenons la solution exacte, ce qui est déjà très important.

Si le modèle est exact, il existe des facteurs F_δ^I permettant une décomposition de tous les W_j. Une matrice W_j s'écrira donc :

$$W_j = \sum_\delta w_j^\delta \; F_\delta^I \; (F_\delta^I)'$$

Ce qui implique que les W_j D admettent ces vecteurs F_δ^I comme base de vecteurs propres, leur somme W.D aussi. Or les composantes que nous calculons sont les vecteurs propres de WD, nous obtenons donc la solution exacte.

Si les données s'éloignent assez peu du modèle, les vecteurs propres des W_j D sont assez proches les uns des autres et la diagonalisation de leur somme en donnera certainement une bonne approximation. Comme nous l'avons déjà dit, ce n'est pas la meilleure approximation au sens du critère ν, mais le critère maximisé étant assez proche de ν, les résultats seront vraisemblablement assez raisonnables. Notons que dans le critère ν, l'inertie des variables d'un groupe intervient par son carré, le choix d'une pondération équilibrant les groupes devient crucial : si un ou deux groupes prédominent ils risquent de déterminer à eux seuls les résultats. Dans le cas où l'équilibre entre les différents groupes est difficile à réaliser, (nombre de variables dans les groupes très différents par exemple), il est peut-être plus raisonnable, a priori, d'utiliser le critère de notre méthode qui rendra compte d'effets plus moyens.

3.4 La méthode dans $(R^I)^*$

Nous nous contentons ici de voir comment se traduit dans l'espace $(R^I)^*$ la représentation simultanée des J nuages N_I^j et celle du nuage moyen.

Rappelons (cf §2.6) que tous les nuages sont représentés dans $(R^I)^*$ par les vecteurs de la base canonique. Ils sont donc confondus, et c'est la métrique définie sur l'espace qui varie avec le groupe de variables. Cette métrique est W_j pour le nuage N_I^j associé au groupe de variable j, $w = \Sigma_j w_j$ pour le nuage associé à toutes les variables et $(1/J)W$ pour le nuage moyen.

Comme toujours, chaque groupe de variables a été surpondéré par un coefficient α_j dont le choix est discuté en 3.5 et qui est inclus ici dans W_j.

Pour représenter simultanément les J nuages N_I^j, nous partons, comme dans les autres cadres, d'une représentation du nuage moyen à laquelle nous superposons des projections des nuages N_I^j.

La représentation du nuage moyen s'impose, c'est la projection sur ses axes principaux d'inertie. Notons $a_1,\ldots,a_\delta$ les vecteurs directeurs de ces axes dans l'ordre décroissant des moments d'inertie. Ces vecteurs sont les vecteurs propres de DW où D joue ici le rôle de l'inertie et W celui de la métrique.

La projection F_δ^I du nuage associé à toutes les variables s'obtient en appliquant W à a_δ, ou en diagonalisant l'opérateur WD. Celle du nuage moyen est $(1/J)\ F_\delta^I$. On a donc :

$$
\begin{aligned}
D\,W\,a_\delta &= \lambda_\delta\, a_\delta \qquad\qquad \text{avec } a'_\delta\, W\, a_\delta = 1\\
F_\delta^I &= Id\ W\ a_\delta = W\ a_\delta\\
a_\delta &= (1/\lambda_\delta)\, D\ F_\delta^I\\
F_\delta^{I\,*} &= (1/J)\ F_\delta^I
\end{aligned}
$$

Nous allons ensuite projeter chaque nuage N_I^j sur ces mêmes axes a_δ. La projection de N_I^j sur a_δ s'écrit puisque W_j est la métrique considérée pour N_I^j :

$$F_{\delta,j}^I = W_j\ a_\delta\ /\ \|\ a_\delta\ \|_j$$

où $\|\ a_\delta\ \|_j$ est la norme de a_δ pour la métrique W_j :

$$// a_\delta //_j^2 = a'_\delta W_j a_\delta$$

$$= (1/\lambda_\delta)^2 F_\delta^I {}' D W_j D F^I$$

d'où :

$$F_{\delta,j}^I = W_j D F_\delta^I / (\lambda_\delta // a_\delta //_j)$$

Nous retrouvons les projections des nuages N_I^j obtenues dans l'espace $\mathbb{R}^I$. Comme dans ce dernier, nous sommes amenés à prendre non pas $F_{\delta j}^I$ mais son produit par $// a_\delta //_j$. En effet, les vecteurs a_δ ne sont généralement pas orthogonaux entre eux. Pour interpréter une représentation plane de N_I^j sur deux axes orthogonaux avec les coordonnées $F_{\delta,j}^I$ et $F_{\delta',j}^I$, comme une projection de ce nuage, il faudrait considérer sur ce plan une métrique différente de l'identité ce qui est illisible. Le cadre $\mathbb{R}^K$ permet de voir qu'en multipliant ces projections par $// a_\delta //_j$ on obtient des projections orthogonales des N_I^j.

De plus, la projection du nuage moyen se trouve alors au centre de gravité des projections des N_I^j.

3-5. Choix des pondérations des groupes

Nous avons introduit dans la méthode proposée des coefficients α_j permettant de pondérer les groupes. Choisir ces α_j, c'est choisir les sur-pondérations de chaque groupe dans le nuage moyen dont on fait l'analyse. C'est aussi choisir la métrique de l'espace $\mathbb{R}^k$ ou les coefficients de la combinaison linéaire de W_j $\quad W = \alpha_1 W_1 + \ldots + \alpha_j W_j$

Nous discutons ici des différents choix possibles. Nous retenons en conclusion deux solutions.

3-5.1. Des α_j égaux

L'avantage de ce choix est la simplicité. Si les α_j sont égaux, le carré des distances dans le nuage moyen est la somme des carrés des distances dans les j nuages, et W est la somme des W_j.

3-5.2. Des α_j donnant la même "importance" à chaque groupe

L'importance des différents groupes dans le nuage moyen peut être très inégale. Elle dépend du nombre de variables qui les composent, de leur poids etc... On peut souhaiter que chaque groupe joue à peu près le même rôle dans ce nuage et chercher à déterminer les α_j en conséquence.

Plusieurs choix sont possibles :

3-5.2.1. Rendre égaux les poids totaux de chaque groupe de variable. Si les variables sont normées et de poids 1, le coefficient α_j seraaaaa inversement proportionnel au nombre de variables du groupe.
Cette solution est simple mais ne possède pas de propriétés particulières intéressantes.

3-5.2.2. Rendre égales les inerties totales de chaque groupe.

Cette solution est équivalente à la précédente si les variables sont normées et de poids 1. Sinon, elle parait plus logique que la précédente car c'est l'inertie des variables et non leur poids qui intervient dans l'analyse.

Cette solution n'est pas forcément satisfaisante. Prenons le cas d'un groupe ayant beaucoup d'axes d'inertie d'importance à peu près égale et d'un groupe ayant un seul axe important correspondant à un nuage presque rectiligne.

Pour simplifier, plaçons-nous dans le cas extrême où les r moments d'inertie du premier groupe sont égaux à 1, r étant grand et où le second a un seul moment d'inertie non nul. L'inertie totale étant la somme des moments d'inertie, si on rend égale l'inertie totale des deux groupes, les moments d'inertie du premier deviendront égaux à 1/ r et celui du second égal à 1.

L'inertie de la projection du premier groupe sur une direction quelconque de l'espace sera toujours inférieure à 1/r alors que celle du second pourra atteindre 1.

Or l'analyse en composantes principales est la recherche de direction de R^I rendant maximum l'inertie des projections des groupes de variables. Il est évident que dans la détermination de la 1ère composante principale, le second groupe aura beaucoup plus d'importance que le premier.

3-5.2.3. Rendre égal le plus grand moment d'inertie de chaque groupe, i.e la plus grande valeur propre des $W_j D$

En effet, comme le montre l'exemple ci-dessus, c'est l'importance de l'inertie dans une direction donnée et non l'inertie totale du groupe qui influe sur la détermination des composantes.

Remarquons que l'on rend ainsi égaux deux nuages homothétiques. En effet, si N_I^j et $N_I^{j'}$ sont homothétiques, les tenseurs associés W_j et $W_{j'}$ sont proportionnels. Pour obtenir des tenseurs dont le plus grand moment d'inertie est égal à 1, nous multiplierons W_j (resp $W_{j'}$) par l'inverse de la plus grande valeur propre de $W_j D$ (resp. $W_{j'} D$). On obtiendra ainsi des tenseurs égaux et donc aussi des nuages identiques.

Remarquons aussi que si l'on remplace un nuage de points par sa projection sur ses premiers axes d'inertie, les pondérations restent inchangées. Ceci n'est vérifié pour aucune autre pondération différente de 1. Or, il est naturel en analyse en composantes principales de considérer le nuage représenté par les premières composantes comme équivalent au nuage associé au tableau. Cette propriété est donc intéressante. Pour nous elle l'est d'autant plus que nous proposerons dans les techniques de mise en oeuvre une variante possible utilisant cette équivalence.

Cette variante permet d'accélérer les calculs au prix d'une certaine approximation des résultats en remplaçant chaque nuage par sa projection sur ses premiers axes d'inertie.

Remarquons encore que dans l'analyse des correspondances multiples qui a fait ses preuves depuis longtemps, c'est cette propriété qui est vérifiée et non les précédentes :
Les valeurs propres des opérateurs de projection associées à chaque groupe de variables indicatrices sont toutes égales, donc a fortiori les plus grandes. Mais les inerties totales ne le sont pas, elles croissent avec le nombre de modalités des variables. Et une variable avec un grand nombre de modalités n'influe pas plus que les autres sur les premiers facteurs, par contre elle influera un plus grand nombre de facteurs.

3-5.2.4. Normaliser les W_j

Puisque choisir les α_j c'est choisir une combinaison linéaire des W_j, il peut sembler naturel pour donner la même importance aux groupes de rendre égales les normes des W_j dans l'espace $(R^I)^2$.
Le carré de la norme de W_j est la somme des carrés des valeurs propres de $W_j D$.

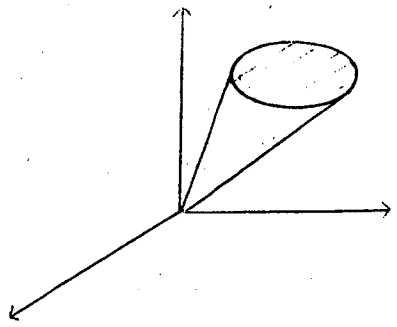
Cette solution présente les mêmes inconvénients que la solution donnant la même inertie à chaque groupe.

3-5.3. Ajuster les W_j

Choisir W étant choisir le tenseur qui résume les tenseurs W_j, on peut tenter de choisir un vecteur de $(\mathbb{R}^I)^2$ qui ajuste au mieux les vecteurs W_j.

Un ajustement classique est l'ajustement aux moindres carrés que l'on obtient en retenant la première composante principale de l'analyse en composante principale des W_j. Cette solution est proposée par ESCOUFIER [cf 6]

Avant de discuter cette solution, rappelons que les W_j ne sont pas des éléments quelconques de $\mathbb{R}^I \otimes \mathbb{R}^I$. Ce sont des tenseurs positifs, i.e. qui s'écrivent $\sum_k m_k U_k \otimes U_k$ avec les m_k positifs. Le produit scalaire de deux tenseurs positifs est toujours positif ou nul, ils ne sont donc jamais de directions opposées et constituent dans un 1/2 cône convexe de faible ouverture de $\mathbb{R}^I \otimes \mathbb{R}^I$.



Les produits scalaires étant positifs, on pourra obtenir une première composante principale de W_j combinaison linéaire à coefficients positifs des W_j (en effet un vecteur propre associé à la plus grande valeur propre d'une matrice à coefficients positifs a toutes ses coordonnées de même signe). Ceci est indispensable pour que les α_j puissent être considérés comme une surpondération des variables et W comme un tenseur d'inertie.

Dans ce calcul, plusieurs solutions se présentent encore, puisque l'on peut affecter des poids aux W_j dans leur analyse en composantes principales. Les deux solutions classiques pour une analyse en composantes principales sont de travailler en normant ou en ne normant pas les W_j : on peut aussi les pondérer comme nous le proposions ci-dessus.

L'ajustement aux moindres carrés nécessite des calculs importants et nous parait, de plus, moins satisfaisant que de prendre tout simplement l'ajustement par la somme.

L'ajustement aux moindres carrés donne un W normé qui maximise :

$$\sum_j < W, W_j >^2$$

Celui par la somme donne un W normé qui maximise :

$$\sum_j < W, W_j >$$

Ce dernier est un bon ajustement dans ce cas particulier où tous les W_j sont de même "sens" et tous les produits scalaires $<W,W_j>$ positifs. L'introduction de carrés (ou de valeurs absolues) est nécessaire lorsque cela n'est pas le cas. De plus, nous avons vu au § 3-5.3 que le produit scalaire dans $(\mathbb{R}^I)^2$ correspondait au carré du produit scalaire dans $\mathbb{R}^I$.

Comparons approximativement les deux ajustements. Si les W_j ont des normes très inégales, le déséquilibre sera renforcé par l'ajustement aux moindres carrés. Dans cet ajustement -beaucoup plus que dans la somme- l'importance des opérateurs de norme élevée sera prépondérante et le rôle des opérateurs de faible norme négligeable. Si les W_j ont des normes proches, ce sont les W_j de direction différente de l'ensemble qui risqueront d'être négligés par l'ajustement aux moindres carrés.

3-5.4. Conclusion

Nous retiendrons deux solutions pour le choix des α_j :

- les α_j tous égaux

On traite alors le tableau de données sans aucune transformation par un programme d'A.C.P classique. C'est la solution la plus simple qui permettra la plupart des calculs.

- les α_j rendant égale à 1 la plus grande valeur propre de chaque groupe. C'est une solution beaucoup plus intéressante qui fait partie intégrante de la méthode proposée.

3.6 Aides à l'interprétation

Les représentations graphiques issues de la méthode que nous venons de décrire comprennent les trois types d'objets auxquels nous nous intéressons : les individus, les variables, les groupes de variables. A propos de chacun de ces types d'objets, il est naturel de se poser les questions usuelles de l'analyse factorielle concernant leurs contributions à la construction des axes et leur qualité de représentation par les axes. Nous indiquons dans les deux premiers paragraphes ce que recouvrent ces notions dans notre cas particulier, en accordant surtout de l'importance aux groupes de variables qui constituent les types d'objets les plus originaux de cette analyse.

En outre, pour aider le dépouillement de la représentation simultanée des intrastructures, il convient de bâtir des critères mesurant la ressemblance entre les intrastructures. C'est ce que nous ferons dans le troisième paragraphe concernant les aides à l'interprétation.

3.6-1 Contributions et qualités de représentation des individus et des variables

Les aides à l'interprétation concernant ces 2 ensembles ne diffèrent pas de celles des analyses factorielles usuelles. Pour chaque individu (et pour chaque variable), on calculera :

- sa contribution à la construction de l'axe δ à l'aide du rapport

$$\frac{\text{Inertie de la projection du point sur l'axe } \delta}{\text{Inertie de l'ensemble des projections sur l'axe } \delta} \times 1000 \ ;$$

- sa qualité de représentation par l'axe δ à l'aide du rapport

$$\frac{\text{Inertie de la projection du point sur l'axe } \delta}{\text{Inertie du point}} \times 1000.$$

.../...

S'il y a des éléments supplémentaires, on calculera uniquement leur qualité de représentation. En particulier, les éléments supplémentaires pourront être :

- des centres de gravité de sous-ensembles d'individus ;
- des composantes principales des analyses partielles.

Les individus considérés seulement du point de vue d'un seul tableau (éléments i^j des nuages N_I^j définis dans R^K (§ 3.1-1) sont traités comme des individus supplémentaires.

3.6-2 Contribution et qualité de représentation des groupes de variables

Selon les cadres de référence dans lesquels on se place, les groupes de variables apparaissent différemment et conduisent à des aides à l'interprétation différentes.

3.6-2 1 Cadre de référence R^K

Les groupes de variables sont représentés par les sous-espaces $R^K{}_j$ de R^K. Les sous-espaces étant orthogonaux deux à deux, l'inertie d'un nuage se décompose additivement suivant les projections de ces nuages sur ces sous-espaces. La contribution du groupe j à l'inertie de la projection du nuage N_I sur un vecteur u de R^K est donc égale au cosinus carré de l'angle θ_j entre u et $R^K{}_j$.

Par ailleurs, un groupe de variables peut être appréhendé à l'aide de la structure qu'il induit sur l'ensemble des individus. Ce point de vue suggère, pour

.../...

apprécier la qualité de représentation du groupe de variables j, de s'intéresser à la qualité de représentation du nuage N_I^j. On est ainsi conduit à calculer le rapport (Inertie du nuage projeté)/(Inertie du nuage total). μ

En outre, nous retrouvons ici un problème déjà soulevé au § 3.1-4. : la projection des éléments des i^j sur un vecteur u de R^K peut être interprétée comme le résultat d'une projection sur le vecteur u_j (composante de u dans $R^K{}_j$) et d'une homothétie de rapport $\frac{1}{\cos \theta_j}$. Il s'ensuit que l'indicateur (Inertie projetée) /(Inertie totale) rend compte de façon pessimiste de la qualité de représentation, en ce sens que la forme d'un nuage N_I^j peut être bien respectée dans la représentation, même si chacun des points est mal représenté (du point de vue de l'indicateur).

Pour prendre en compte ce phénomène, on peut penser diviser les indicateurs de qualité par $\cos^2\theta_j$. Mais ces indicateurs perdraient ainsi leur propriété d'additivité axe par axe ; en particulier, leur somme, pour un même individu sur plusieurs axes, pourrait dépasser 1.

3.6-2 2 Cadre de référence R^I

Les groupes de variables sont représentés par les sous-nuages N_K^j, regroupant les vecteurs associés aux variables du groupe j. (cf § 3.2-1). Ce point de vue suggère de s'intéresser aux contributions et qualités de représentation de N_K^j.

La contribution se définit par la somme des contributions des variables du groupe j. Cet indicateur se confond avec la quantité $(\cos \theta_j)^2$ envisagée à l'alinéa précédent. (cf § 3-1-7).

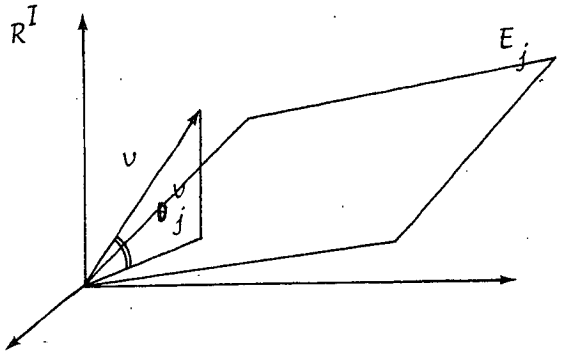
.../...

La qualité de représentation d'un groupe j peut être mesurée, dans R^I, à l'aide de plusieurs critères.

a) L'inertie de la projection de N_K^j rapportée à l'inertie totale de ce nuage. Si le groupe j ne comprend qu'une variable, ce critère est l'indicateur usuel de qualité de représentation.

b) Le cosinus carré de l'angle formé par l'axe factoriel et le sous-espace engendré par le groupe j.

Notation :



v vecteur unitaire porté par l'axe factoriel

E_j sous-espace engendré par le groupe j.

θ_j^v angle entre v et E_j

Dans le cas (général) où les variables sont centrées réduites, $\cos^2\theta_j^v$ est le carré du coefficient de correlation multiple entre la "variable" v et le groupe j. Comme précédemment, il est équivalent à l'indicateur usuel lorsque le groupe j ne contient qu'une seule variable.

Toutefois, ce critère ne tient compte du groupe de variables n° j qu'au travers du seul sous-espace E_j : nous avons déjà critiqué ce point de vue (§ 2.5-3 b).

c) La qualité de représentation des composantes principales issues de l'ACP (partielle) du nuage des j variables.

Nous avons déjà cité (§ 3.2.4.)l'intérêt de projeter "en éléments supplémentaires", ces composantes principales. Si celles-ci ont pu être clairement interprétées, ce troisième type d'indicateur fournit un point de vue intéressant : il indique quels aspects des sous-nuages sont bien représentés, de manière plus synthétique qu'en procédant variable par variable. Nous retiendrons, dans l'analyse, les indicateurs a) et c).

Il n'est pas possible de comparer, a priori, les valeurs prises par le critère a) et son homologue défini dans R^K. En effet, en appelant :

$I(j)$ l'inertie totale du nuage associé au groupe j. Cette inertie est la même, que l'on considère le nuage N_I^j dans R^K ou N_K^j dans R^I.

$I_u(N_I^j)$ l'inertie du nuage N_I^j selon la direction u de R^K ;

$I_v(N_K^j)$ l'inertie du nuage N_K^j selon la direction v de R^I ;

alors, les qualités de représentation du groupe j par u dans R^K (notée $Q_K^u(j)$ et par v dans R^I(notée $Q_I^v(j)$) s'écrivent :

$$Q_K^u(j) = \frac{I_u\ N(j_I)}{I_j} \qquad Q_I^v(j) = \frac{I_v\ (N_K^j)}{I_j}$$

Les dénominateurs de ces deux quantités sont les mêmes. Les numérateurs ne sont égaux que si u et v sont vecteurs propres homologues associés aux nuages N_I^j et N_K^j. Naturellement, il n'y a pas de relation d'ordre a priori entre ces deux numérateurs.

3.6-2 3 Cadre de référence R^{I^2}

Dans ce cadre, les groupes de variables sont représentés par des vecteurs. Les aides à l'interprétation pourront être définies à l'aide des critères usuels.

La contribution d'un groupe de variables à un axe mesure le rôle joué par ce groupe dans la construction des axes factoriels. Dans R^K, le critère optimisé par les axes factoriels est la somme des projections. La contribution d'un élément pour un axe factoriel sera donc la coordonnée de cet élément sur cet axe. Comme toujours, cette quantité sera rapportée à la valeur du critère pour l'axe, c'est-à-dire ici la somme des projections. Nous retrouvons ici le même critère suggéré par les cadres R^K et R^I.

Par contre, la notion de qualité de représentation est indépendante du critère que l'on cherche à optimiser. Nous utiliserons donc, pour la mesurer, le critère usuel (Inertie projetée)/(Inertie totale), appliquée aux vecteurs W_j^0.

Ici encore, il n'est pas possible de comparer, a priori, les valeurs prises par ce critère et par ses homologues définis dans les autres cadres de référence. En effet, en considérant les notations du paragraphe précédent, et en les complétant par :

$I(j) = \sum_{\delta} \lambda_{\delta}^{j}$ L'inertie totale du nuage N_K^j dans R^I peut s'exprimer en fonction de la somme des valeurs propres λ_{δ}^{j} de l'opérateur $W_j D_p$ associé.

$Q_{I^2}^{w}(j)$ La qualité de représentation du groupe j par l'axe w de R^{I^2}.

Si l'on choisit v et w tels que $w = v \otimes v$, on a :

$$Q_I^v(j) = \frac{I_v(N_K^j)}{\sum_{\delta} \lambda_{\delta}^{j}} \quad \text{et} \quad Q_{I^2}^{w}(j) = \frac{[I_v(N_K^j)]^2}{\sum_{\delta} (\lambda_{\delta}^{j})^2}$$

Il n'y a pas de relation d'ordre a priori entre ces deux quantités.

Remarque

Quel que soit le cadre de référence adopté, la notion de contribution d'un groupe de variables conduit au même calcul. Cela n'est pas tout à fait étonnant puisque les critères optimisés dans chacun des cadres constituent des interprétations différentes d'un même critère.

Par contre, la qualité de représentation d'un groupe de variables diffère selon le cadre dans lequel on l'exprime. Cette particularité n'est pas obligatoirement surprenante puisque les objets qui représentent les groupes de variables sont de types différents: point dans R^{I^2}, nuage de I ou K_j points dans R^K ou R^I.

3.6-2 4 Résumé

Les aides à l'interprétation concernant les groupes de variables sont, pour chaque groupe et chaque axe (noté u, v, w dans R^K, R^I, R^{I^2}) :

aïde	Interprétation		
	dans R^K	dans R^I	dans R^{I2}
contribution	$\cos^2 \theta_j^u$	Σ contributions des variables du groupe j à l'axe ν.	$\dfrac{\text{Projection de } W_j D \text{ sur } w}{\sum_j (\text{projection des } W_j D \text{ sur } w)}$
qualité	$Q_K^u(j) = \dfrac{I\ (N_I^j)}{\sum_\delta \lambda_\delta^j}$	$Q_I^\nu(j) = \dfrac{I_\nu(N_K^j)}{\sum_\delta \lambda_\delta^j}$	$Q_{I^2}^w(j) = \dfrac{[I_\nu(N_K^j)]^2}{\sum_\delta (\lambda_\delta^j)^2}$

Les trois interprétations de la contribution conduisent au même calcul. Les trois qualités constituent des indicateurs différents.

3.6-3 Critère de ressemblance entre les intrastructures

La ressemblance entre les intra-structures peut être appréhendée de différentes façons selon les cadres de référence dans lesquels on se situe. Nous abordons successivement ces différents points de vue.

En réalité, les critères que nous allons présenter peuvent être introduits indépendamment des cadres de référence. Ces derniers servent ici simplement à faciliter l'exposé.

3.6-3 1 Cadre de référence R^K

Dans cet espace, nous avons cherché une représentation simultanée qui optimise un critère réalisant lui-même un compromis entre 2 propriétés (cf § 3.1-2).

(p1) Chacun des nuages N_I^j est bien représenté

(p2) Les points homologues i^j $(j=1,J)$ sont le plus rapprochés possible.

.../...

Nous avons vu que chacune de ces 2 propriétés conduit à optimiser un critère exprimable en terme d'inertie. Soit, avec les notations des § 3.1-2 et 3:1-3, dans lesquels on introduit la partition P de $\bigcup_j N_I^j$ dont les classes sont les ensembles $\{i^j \mid j=1,J\}$ d'éléments homologues.

Pour $(p1)$, le critère associé est C1 = inertie totale de $\bigcup_j N_I^j$

Pour $(p2)$, le critère associé est C2 = inertie intraclasse de $\bigcup_j N_I^j$ muni de la partition P.

.Le compromis entre ces 2 propriétés a été réalisé en optimisant le critère global : C = C1-C2 = Inertie totale - Inertie intra = Inertie Inter

Pour mesurer la ressemblance entre les intra-structures telles qu'elles apparaissent dans la représentation simultanée, cette démarche nous conduit à calculer, pour chaque axe factoriel, le rapport (Inertie inter)(Inertie totale). Cet indicateur est toujours compris entre 0 et 1 et vaut :

1, si les intra-structures sont parfaitement confondues dans la direction étudiée ;

0, si l'inertie du nuage moyen est nulle. De telles directions sont dues exclusivement a l'aspect articifiel de R^K. En effet, seules les directions de cet espace selon lesquelles l'inertie du nuage moyen est non nulle sont intéressantes.

Cette dernière situation appelle une remarque. Sauf situation triviale, elle ne peut se produire pour la direction principale d'inertie du nuage moyen. En effet, cela signifierait que l'inertie du nuage moyen est nulle, ce qui ne se produit que si toutes les coordonnées de tous les points sont nulles.

Nous avons ici défini un critère associé à un axe. L'adaptation de cette définition à un sous-espace de plus grande dimension ne pose aucun problème. Natu-

.../...

rellement, la valeur obtenue pour un plan n'est pas égale à la somme des valeurs pour chacun des axes, mais à leur moyenne pondérée par la part de l'inertie de $\bigcup_j N_I^j$ le long des axes. Ainsi, en appelant :

$Cl(\ell)$ l'inertie totale de $\bigcup_j N_I^j$ dans la direction de l'axe ℓ ;

$Cl(\ell+m)$ l'inertie totale de $\bigcup_j N_I^j$ dans la direction du plan $(\ell+m)$ défini par les axes ℓ et m.

On a :

$$Cl(\ell+m) = Cl(\ell) + Cl(m)$$

$$C(\ell+m) = C(\ell) + C(m)$$

d'où :

$$\frac{C(\ell+m)}{Cl(\ell+m)} = \frac{Cl(\ell)}{Cl(\ell) + Cl(m)} \frac{C(\ell)}{Cl(\ell)} + \frac{Cl(m)}{Cl(\ell) + Cl(m)} \frac{C(m)}{Cl(m)}$$

critère pour le plan $(\ell+m)$ — critère pour l'axe ℓ — critère pour l'axe m

Inertie de $\bigcup_j N_I^j$ selon ℓ rapportée à celle selon $(\ell+m)$

Inertie de $\bigcup_j N_I^j$ selon m rapportée à celle selon $(\ell+m)$

Nous venons de décrire un critère mesurant globalement la ressemblance entre les intra-structures. Il est possible d'adapter cette définition à 2 intra-structures, l'une des deux pouvant être l'intra-structure moyenne. Le calcul de ces quantités apporte des éléments de réponse à des questions du type :

- Ces deux représentations d'intra-structure se ressemblent t-elles? Quelles sont les représentations qui se ressemblent le plus?

- Cette représentation d'intra-structure ressemble t-elle à la représentation moyenne? Quelles sont les représentations les plus proches de la représentation moyenne?

.../...

3.6-3 2 Cadre de référence R^I

Dans ce cadre, la recherche de la représentation simultanée a été formulée en termes de recherche d'un ensemble de vecteurs (cf § 3.2-4). Nous avions noté v_j le vecteur correspondant à la représentation des individus considérés du point de vue du groupe de variables j.

Dans cet espace, les propriétés (*p1*) et (*p2*) ont été traduits de la façon suivante (§ 3.2-3) :

(*p1*) La norme des v_j est grande.

(*p2*) Les v_j sont bien corrélés entre eux.

La ressemblance, du point de vue d'un axe, entre les intra-structures associées à deux groupes j et j' sera mesurée par le coefficient de corrélation entre v_j et $v_{j'}$ (noté $r(v_j, v_{j'})$). De même, la ressemblance entre l'intra-structure associée au groupe j et le compromis sera mesurée par le coefficient de corrélation entre v_j et v (noté $r(v, v_j)$).

Pour mesurer la ressemblance globale entre l'ensemble des intra-structures, on se heurte au problème classique de la mesure de la liaison d'un groupe de variables. On utilisera ici :

soit $\frac{1}{J^2} \sum_j \sum_\ell r(v_j, v_\ell)$ — critère déjà proposé par HORST [11] pour mesurer la liaison entre plusieurs variables

soit $\frac{1}{J} \sum_j r(v_j, v)$ — qui correspond mieux à la démarche de l'analyse multicanonique.

Enfin, si l'on cherche des critères de ressemblances s'appuyant sur deux

.../...

axes, on utilisera la moyenne des critères appliqués à chacun des axes.

Remarque :

On peut songer, dans les coefficients de corrélation moyens proposés, à pondérer chaque coefficient de corrélation par la qualité de représentation des intra-structures mises en cause. On obtient ainsi un critère très proche de celui introduit à partir de l'espace R^K. Ceci est dû au résultat suivant :

En appelant

$\tilde{X}_j$ le tableau centré associé au groupe j, complété par des 0 pour être de la taille de X. On a la relation $X = \sum_j \tilde{X}_j$

u une composante principale de X.

On a : Inertie du compromis selon u (inertie inter classe du § précédent) :

$$\frac{1}{J^2}\, u'MX'DXMu$$

$$= u'M(\sum_j \tilde{X}_j')\ D(\sum_\ell \tilde{X}_\ell)Mu = \sum_j \sum_\ell u'M\,\tilde{X}_j'\,D\,\tilde{X}_\ell\,Mu = \sum_j \sum_\ell \langle v_j, v_\ell \rangle_M$$

$$= \sum_j \sum_\ell \| v_j \| \, \| v_\ell \| \, r(v_j, v_\ell)$$

En n'appliquant qu'une fois $X = \sum_j \tilde{X}_j$, on obtient :

$$\frac{1}{J^2}\, u'MX'DXM_u = \sum_j \| v_j \| \, \|v\| \, r(v_j, v_\ell)$$

Ces résultats expriment l'inertie inter-classe du paragraphe précédent, (et par conséquent le critère C/C1) en termes de sommes pondérées des coefficients de corrélation.

.../...

3.6-3 3 La multiplicité des critères de ressemblance entre intra-structures

Nous venons d'introduire un grand nombre de critères pour apprécier la ressemblance entre intra-structures. En outre, cette présentation n'est pas exhaustive et on peut en imaginer beaucoup d'autres. Les origines de cette multiplicité sont diverses : à ce stade de l'exposé, il peut être utile de les énumérer afin de dégager quelques fils directeurs utiles pour appréhender ce vaste ensemble.

a) Deux représentations d'intra-structures homothétiques doivent-elles être considérées comme semblables ?

Si la réponse est oui, on est conduit à examiner uniquement le coefficient de corrélation. Sinon, on prendra en compte les inerties.

Bien entendu, il n'est pas facile de répondre formellement à cette question. Compte tenu de la méthode, ce type de résultats ne doit pas se produire.

b) La comparaison de plusieurs intra-structures doit-elle s'appuyer sur les comparaisons 2 à 2 ou sur les comparaisons entre les intra-structures et le compromis ?

c) Comment agréger, pour un groupe d'intra-structures, des critères de comparaison établis pour deux intra-structures ?

Quelle que soit la réponse à b), on aggrège des quantités de même type.

Remarquons que la réponse à la question a) donne des éléments (à défaut d'impliquer logiquement) pour répondre ici. Si on utilise le rapport d'inerties (réponse non à la question a)), ce dernier se définit de la même manière pour comparer deux ou plus de deux intra-structures. Si on utilise les coefficients de corrélation (réponse non à la question a)), il semble peu cohérent de les pondérer par des inerties (coefficients de pondération les plus "intuitifs") dans une aggrégation.

d) Comment agréger, pour un sous-espace de dimension supérieure à 1,

.../...

des critères définis pour chacune des dimensions de ce sous-espace ?

Ici encore, cette question n'est vraiment embarassante que si l'on répond oui à la question a).

3.6-3 4 Que choisir ?

Il est difficile de trancher actuellement en faveur de certains critères. Il est donc prudent de calculer l'ensemble de ces critères, c'est-à-dire, pour chaque axe et pour les plans représentés graphiquement.

1) Critères fondés sur un rapport d'inertie :

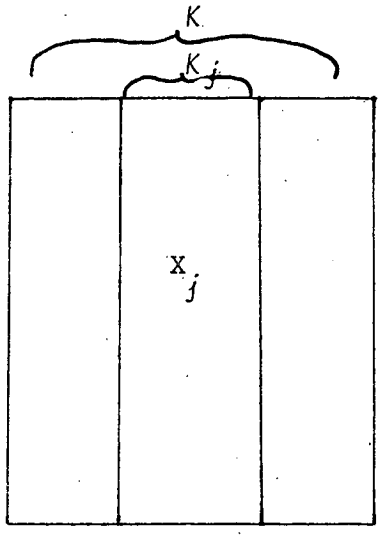
- critères concernant les intra-structures prises 2 à 2 ;
- critères concernant chaque intra-structure et l'intra-structure moyenne ;
- critère global.

2) Critères fondés sur des coefficients de corrélation

- idem.

3.7 Résumé

Les données - J groupes de variables sur le même ensemble d'individus I.



$K = \bigcup_j K_j$

v_k variable

m_k poids de v_k

Le tableau de données X divisé en J sous tableaux $X_1, \ldots, X_J$

La pondération des groupes

Le poids des variables du groupe j est multiplié par un coefficient α_j. Plusieurs pondérations sont possibles, la plus intéressante est l'inverse de la première valeur propre à l'A.C.P. du groupe X_j. Pour chaque groupe, l'inertie maximum dans une direction vaudra alors 1.

Analyse en composantes principales du tableau X surpondéré

= composantes principales F_δ^I et $F^{*I}_\delta = (1/J)\ F_\delta^I$

= variables générales de l'analyse des liaisons des groupes de variables (recherche de variables orthogonales deux à deux les plus liées à tous les groupes).

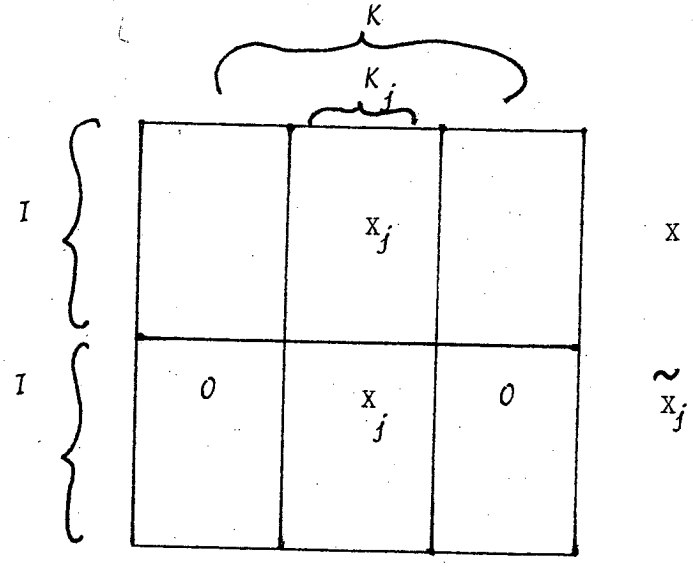
= projection sur ses axes principaux d'inertie d'un nuage compromis entre les nuages d'individus N_I^j définis par chaque groupe de variables.

= facteurs du modèle INDSCAL.

Projection des composantes principales en éléments supplémentaires

Permet de comparer facilement ces composantes qui sont projetées sur un espace ajusté au sens des moindres carrés.

Projection en individus supplémentaires des j tableaux $\tilde{X}_j$ (X_j complétée par des zéros)



= projection de $\tilde{X}_j$ sur l'axe s : $F^I_{s,j} = 1/\lambda_s \; W_j \; D \; F^I_s$

= représentation simultanée des J nuages d'individus N^j_I associés à chaque groupe de variables. Ces représentations simultanées sont des projections de ces nuages dans des directions d'inertie importante et se ressemblant entre elles.

= le nuage moyen est au centre de gravité de ces J nuages.
$F^{*I}_s(i) = 1/J \sum_j F^I_{s,j}$

= variables canoniques de l'analyse des liaisons : combinaison linéaire des variables du groupe j qui est liée à F^I_s et qui extrait une part importante de l'inertie du groupe .

Inertie de la projection des variables du groupe j sur F^I_s

= contribution du groupe de variables à l'inertie du facteur d'ordre s dans $\mathbb{R}^I$, dans $\mathbb{R}^K$ et dans $(\mathbb{R}^I)^2$.

= liaison entre la variable F^I_s et le groupe j (c'est la somme de ces liaisons qui est maximisée dans l'analyse des liaisons).

= coordonnée de la projection dans $(\mathbb{R}^I)^2$ du tenseur W_j associé au groupe j sur le tenseur $F^I_s \otimes F^I_s$. Ceci permet de comparer globalement les groupes de variables, car le produit scalaire entre les tenseurs W_j et $W_{j'}$ mesure la liaison entre les deux groupes de variables. La projection des W_j sur des espaces de petites dimensions permet de les comparer rapidement.

= poids affecté au facteur F^I_s par le groupe j dans le modèle INDSCAL.

Aides à l'interprétation : Ressemblance entre les projections des N_I^j

* Corrélation entre F_δ^I et $F_{\delta,j}^I$

Mesure la ressemblance entre les projections normées du nuage moyen et du nuage N_I^j sur l'axe δ. Mesure l'adéquation de chaque groupe de variables au modèle INDSCAL. (Si ces corrélations valent 1 pour tous les axes, le modèle est vérifié pour le groupe j).

* $\quad \| F^*{}_\delta^I \|^2 / (1/J) \sum_j \| F_{\delta,j}^I \|^2$

C'est sur l'axe s l'inertie du nuage moyen N_I^* (inertie intra) divisée par l'inertie de tous les nuages N_I^* (inertie totale). Plus ce rapport est proche de 1, plus l'inertie intra est faible et plus les projections F_1^I des différents nuages se ressemblent.

Les corrélations et les rapports d'inertie sont aussi calculés pour tous les couples de facteurs.

* Inertie "intra" de chaque individu i. C'est l'inertie des projections de i dans les différents nuages, $F_{\delta,j}^I(i)$, autour du centre de gravité $F_\delta^{*I}(i)$. Elle permet de repérer les individus dont la position varie dans les différents groupes de variables.

Aides à l'interprétation. Qualité de représentation du groupe j

* Qualité de représentation des variables du groupe dans $\mathbb{R}^I$.
* Qualité de représentation du nuage N_I^j associé au groupe j dans $\mathbb{R}^K$.
* Qualité de représentation des tenseurs associés W_j (dans $(\mathbb{R}^I)^2$). (qualité de l'approximation des produits scalaires par le modèle INDSCAL). Ces qualités de représentation s'ajoutent sur les différents axes.
* Liaison entre F_δ^I et le groupe j (cette liaison est comprise entre 0 et 1 et atteint 1 lorsque F_δ^I est une direction d'inertie maximum du groupe.

3.8 Individus, variables et groupe de variables supplémentaires

3.8.1 - Individus supplémentaires

Comme dans toute analyse en composantes principales, des individus pourront être mis en éléments supplémentaires, avec un poids nul. Ces individus n'influeront pas sur les résultats concernant les individus principaux, on calcule simplement la projection de leur représentation dans le nuage N_I^*

et dans les différents nuages N_I^j .

Dès que le nombre d'individus est assez grand, la lecture des graphiques des représentations simultanées devient très complexe. En effet, le nombre de points concernant les individus seulement, est égal à $card\ I \times (card\ K+1)$. (i.e. nombre d'individus multiplié par le nombre de groupe de variables, plus un pour le nuage moyen). La lecture des aides à l'interprétation (stabilité de chaque individu, ressemblance globale entre la représentation d'un N_I^j et la représentation moyenne etc...) facilite beaucoup le dépouillement. Mais, il n'en restera pas moins nécessaire, souvent, de remplacer l'étude de chaque individu par l'étude de classes d'individus ayant un caractère commun. Pour celà, on mettra en éléments supplémentaires les centres de gravité de ces classes.

3.8.2-Variables supplémentaires

Les variables supplémentaires sont traitées comme en A.C.P. : projetées sur les axes de l'analyse ;leur qualité de représentation est calculée.

3.8.3 - Groupe de variables supplémentaires

Tout un groupe de variables peut être mis en élément supplémentaire. Si ce groupe est homogène, il peut être intéressant de le comparer aux autres groupes avec tous les moyens mis en oeuvre pour ces derniers sans qu'il ait influé sur le nuage moyen et les résultats de l'analyse.

Voyons si tous les calculs effectués sur les groupes principaux peuvent s'appliquer à un groupe supplémentaire.

* La normalisation du nuage N_I^j

Si on veut comparer le nuage associé au groupe supplémentaire aux autres nuages, il faudra le normaliser de la même façon en surpondérant les variables du groupe, ce qui nécessite l'analyse du tableau supplémentaire.

* La projection des composantes principales du groupe

Elle ne pose aucun problème et permet de comparer la forme générale du nuage de variable avec celle du nuage moyen et des autres nuages.

* Les variables canoniques du groupe

Dans l'analyse des liaisons, les variables générales sont calculées sans tenir compte du groupe supplémentaire; on peut cependant chercher les variables de ce groupe les plus"liées" à ces variables générales. Le critère chosi et sa solution $F^I_{\delta,j} = (1/\lambda_\delta)\ W_j\ D\ F^I_\delta$ s'appliquent sans aucun changement à des groupes supplémentaires.

* La projection simultanée du nuage N_I^j

Pour les groupes de variables non supplémentaires, les fonctions $F^I_{\delta,j}$ s'interprétaient aussi dans $\mathbb{R}^K$ comme des projections orthogonales du nuage N_I^j sur des espaces de petite dimension. Ici quelques difficultés apparaissent. En effet, l'ensemble K_j des variables supplémentaires n'est pas contenu dans l'ensemble K des variables actives, et $\mathbb{R}^{Kj}$ n'est donc pas contenu dans l'espace $\mathbb{R}^K$ où le nuage moyen a été construit.

On pourrait, bien sûr, considérer l'espace $\mathbb{R}^{KUKj}$. Le nuage N_I^j y serait représenté comme les nuages actifs dans le sous espace $\mathbb{R}^{Kj}$. Mais, le nuage moyen N_I^* étant construit en considérant seulement les groupes de variables actifs, (ou en donnant un poids nul aux variables du groupe supplémentaire), ce nuage N_I^* est contenu dans l'orthogonal du sous espace $\mathbb{R}^{Kj}$. Ses axes d'inertie le seront aussi, les projections de N_I^j sur ces axes vaudront donc 0. C'est ce que l'on obtiendrait d'ailleurs en procédant au même calcul que pour les autres tableaux, i.e., en mettant en élément supplémentaire le tableau $\tilde{X}_j$, (le tableau X_j complété par des zéros).

Cependant, on peut tenter d'interpréter $F^I_{\delta,j}$ comme une projection du nuage N_I^j dans cet espace $\mathbb{R}^{KUKj}$. Sur cet espace on considère la métrique diagonale des poids M (les variables du groupe supplémentaire étant surpondérées par le coefficient de normalisation du nuage). La fonction $F^I_{\delta,j}$ est alors homothétique à la projection de N_I^j sur l'axe $X' \, D \, F^I_\delta$.
En effet (cf §2-4-1) :

$$\text{Projection de } N_I^j \text{ sur } X' \, D \, F^I_\delta = \tilde{X}_j \, M \, X \, D \, F^I_\delta \, / \, \| X \, D \, F_\delta \|$$

$$= W_j \, D \, F^I_\delta \, / \| X \, D \, F_\delta \|$$

Si le tableau X tout entier était intervenu pour le calcul des composantes, on aurait (cf 2-4-1) $// X \, D \, F_\delta // = \lambda_\delta$, et $F^I_{j,\delta}$ serait exactement la projection de N_I^j sur $X' \, D \, F^I_\delta$. Mais ici, ce n'est pas le cas, $F^I_{j,\delta}$ n'est que l'homothétique de cette projection. Le problème se complique encore lorsque l'on considère plusieurs axes, car rien ne dit que les axes $X \, D \, F_\delta$ et $X \, D \, F_{\delta'}$ seront orthogonaux dans l'espace $\mathbb{R}^{KUKj}$. Les fonctions $F^I_{\delta,j}$ ne permettent donc pas d'obtenir des projections orthogonales des N_I^j sur des sous espaces de dimension supérieure à un. Elles ne présentent donc guère d'intérêt.

Parvenus à cette conclusion, il est permis de se demander si, a priori, une représentation simultanée du nuage associé à un groupe de variables supplémentaires pourrait avoir un intérêt. La question préalable est de savoir pourquoi ce groupe a été mis en élément supplémentaire. Or, mettre un groupe en élément supplémentaire parait utile essentiellement lorsque ce groupe rend hétérogène l'ensem-

ble des groupes,i.e. lorsqu'il est relativement différent des autres ; ou bien lorsqu'il peut expliquer les facteurs. Dans un cas comme dans l'autre, une représentation simultanée ne serait guère utile. Dans le cas d'un groupe assez différent des autres les indices globaux, projections des composantes principales, etc... permettent de mesurer et préciser ces différences, mais superposer le nuage N_I^j à des nuages qui ne lui ressemblent pas assez,n'a pas d'intérêt. Lorsque le groupe supplémentaire est utilisé comme élément explicatif, on mesurera globalement l'influence de ce groupe de variables sur les autres, mais on ne s'intéressera pas à chaque individu.

Inertie de la projection des variables du groupe sur F_δ^I

Cet indice ne peut s'interpréter comme une contribution à l'inertie du facteur F_δ^I .Mais il reste tous les autres aspects : mesure de la liaison entre F_δ^I et le groupe, coordonnée de W_j sur le tenseur associé à F_δ^I dans $(\mathbb{R}^I)^2$ et poids affecté à ce facteur par le groupe j dans INDSCAL.

Conclusion

Pour les groupes de variables supplémentaires, nous calculerons tous les indices obtenus pour les groupes actifs mais nous ne chercherons pas de représentation simultanée du nuage d'individus associé à ce groupe.

4 - COMPARAISON AVEC D'AUTRES METHODES

Dans ce paragraphe nous comparons la méthode proposée avec chacune des méthodes évoquées dans l'introduction.

Suivant le cas, on choisira seulement l'un ou l'autre aspect de notre méthode puisqu'aucune autre méthode ne les présente simultanément.

4.1 - Les analyses canoniques généralisées

Nous avons déjà évoqué ces techniques dans les objectifs exprimés en terme de liaison entre groupes de variables (§ 1-6) et dans leur étude (§ 1-3).

La comparaison avec l'analyse multicanonique au sens de CARROLL [cf 5] est exposée dans ce dernier paragraphe puisque nous avons cherché une méthode basée sur le même principe et se confondant avec elle dans le cas des variables qualitatives.

Résumons simplement les résultats :
Le principe de ces techniques est le suivant : la recherche d'une suite de variables générales orthogonales deux à deux rendant maximum la somme de leurs liaisons avec tous les groupes de variables; puis, pour chaque variable générale et pour chaque groupe, la recherche d'une combinaison linéaire des variables du groupe liée à la variable générale.

C'est la mesure des liaisons entre une variable et un groupe de variables qui diffère :

Dans l'analyse multicanonique au sens de CARROLL, c'est le carré du coefficient de corrélation multiple qui ne tient compte que du sous-espace engendré par les variables.

Dans notre méthode, c'est un coefficient qui tient compte de l'inertie des variables du groupe dans les différentes directions de l'espace.

Les variables générales que nous obtenons seront donc beaucoup plus attirées par les directions de grande inertie que par les directions de faible inertie. La variance expliquée par ces variables sera donc plus grande. Les résultats sont, d'autre part, beaucoup plus stables que ceux de l'analyse multicanonique classique qui sont très sensibles à des petites variations des données. (le sous-espace engendré par les variables est beaucoup moins stable que l'opérateur d'inertie (cf § 2-5.1.c).

Les variables canoniques sont, dans l'analyse multicanonique, les projections des variables générales et dans notre méthode leur image par $W_j D$.

Les premières sont plus correlées que les secondes aux variables générales, mais par contre moins représentatives des groupes, au sens de la variance expliquée. Nous augmentons donc la variance expliquée, à la fois par le choix de la variable générale et par le calcul ultérieur des variables de chaque groupe.

Les résultats sont donc un peu différents, plus stables et plus représentatifs des groupes dans notre méthode que dans l'analyse classique. L'étude du cas trivial où tous les groupes sont identiques (cf §3.2.3)semble bien montrer aussi que du simple point de vue de l'analyse des liaisons, la méthode proposée est une amélioration de l'analyse classique.

Mais bien entendu, l'avantage primordial de notre méthode est le fait que ses résultats s'interprètent, en plus, comme une analyse en composantes principales classique et comme une représentation simultanée des nuages d'individus définis par tous les groupes de variables, et d'un nuage compromis entre eux.

Cas de deux groupes de variables

On peut rapprocher cette technique de celle proposée par WOLLENBERG [cf 21] comme une alternative de l'analyse canonique.Remarquant que dans l'analyse canonique, la variance expliquée par les variables obtenues risque d'être faible, il propose de chercher à maximiser non plus la corrélation mais la "redondance". Il cherche donc une suite de variables orthogonales deux à deux, combinaisons linéaires de variables du premier groupe, maximisant la somme des carrés des corrélations avec les variables du second groupe. On reconnait ici un critère tout à fait analogue au nôtre (nous donnons seulement en plus la possibilité de pondérer variables et individus). Mais, ce critère est appliqué, comme en analyse canonique, directement entre les deux groupes sans introduction de variables générales.

La solution est donnée par la diagonalisation d'un opérateur qui, dans nos notations s'écrirait $P_1 W_2 D$. Le problème est posé de manière symétrique pour les deux groupes, mais il n'y a plus de symétrie des résultats comme en analyse canonique :
les valeurs propres de $P_1 W_2 D$ et $P_2 W_1 D$ sont différentes et il n'y a pas de relation simple entre leurs vecteurs propres, donc pas de relation entre les deux suites de variables obtenues pour le premier groupe et pour le second groupe.

Notre méthode apparait comme une généralisation de celle de WOLLENBERG tout à fait parallèle à la généralisation de l'analyse canonique proposée par CARROLL Cependant, il ne s'agit pas exactement d'une généralisation, car dans le cas de deux groupes seulement, ces deux méthodes ne se confondent pas. La méthode que nous proposons rétablit la symétrie entre les deux groupes de variables,et les deux suites de variables "canoniques" sont en relation l'une avec l'autre. Nous cherchons comme en analyse canonique, des couples de variables liées entre elles, l'une étant une combinaison linéaire du premier groupe et l'autre du second, en passant techniquement par l'intermédiaire d'une variable générale. Dans la méthode de WOLLENBERG, on cherche indépendamment deux suites de variables combinaisons linéaires des variables d'un groupe, qui soient liées à l'autre groupe.

Cette méthode parait donc intéressante aussi dans le cas de deux groupes de variables seulement, où elle présente l'avantage par rapport à l'analyse canonique de fournir des variables expliquant mieux la variance des groupes.

Cependant, une des propriétés de l'analyse canonique est perdue. C'est l'orthogonalité des variables canoniques d'un même groupe. Cette propriété n'était pas vérifiée pour l'analyse canonique généralisée de CARROLL lorsque le nombre de groupes était supérieur à deux.

Mais cette propriété était-elle vraiment importante ?

Et n'y a t-on pas gagné en la remplaçant par la suivante :

Les variables canoniques sont des projections orthogonales des nuages d'individus associés à chaque groupe de variables sur des axes orthogonaux deux à deux. Ceci permet une représentation simultanée des nuages que ne fournit pas l'analyse canonique classique.

4.2 - Analyse en composantes principales des opérateurs

Nous avons déjà évoqué plusieurs fois cette méthode et esquissé sa comparaison avec la méthode proposée ici.

Les groupes de variables sont représentés dans l'espace euclidien $(R^I)^?$ par des vecteurs. Le produit scalaire de ces vecteurs ou le cosinus de leur angle mesure leur liaison. Pour obtenir une visualisation simple de ces liaisons, il est suggéré [cf 3 et 6] de projeter ces vecteurs sur un espace de petite dimension en les ajustant au mieux. La solution est une analyse en composantes principales dont les résultats s'obtiennent en diagonalisant la matrice de leurs produits scalaires. C'est ce que nous appelons ici l'analyse des opérateurs.

Pour visualiser les distances entre opérateurs plutôt que leur liaison, il suffit de faire une analyse analogue en prenant comme origine les centres de gravité de leur nuage. Les opérateurs sont alors traités comme le sont les individus dans l'A.C.P. classique.

Dans la méthode que nous proposons, le volet ayant $(R^I)^2$ comme espace de référence peut se comparer à ces analyses des opérateurs. Les représentations des groupes de variables sont isomorphes et dans les deux cas, elles sont projetées dans un espace de petite dimension.

L'ajustement du sous espace est meilleur dans l'analyse des opérateurs, les groupes de variables sont donc mieux représentés. Mais, dans notre méthode les composantes ont une signification, individus et variables apparaissent sur des graphiques annexes ce qui permet une interprétation fine des résultats, alors que l'analyse des opérateurs ne permet de conclure que sur la plus ou moins grande proximité des opérateurs entre eux sans aucun élément d'interprétation.

Et ainsi que le disent les auteurs dans l'article précité [cf 3]: "Pour obtenir les résultats les plus satisfaisants, il faut effectuer son analyse de façon à réaliser un harmonieux compromis entre "qualité de représentation" d'une part et "facilité d'interprétation" d'autre part.

4.3 - La méthode STATIS et ses variantes

La méthode STATIS proposée et programmée par H. L'HERMIER DES PLANTES s'applique à la comparaison de tableaux de type plus général que ceux auxquels nous nous sommes volontairement restreints.

Cette méthode et les méthodes très proches proposées par Y. ESCOUFIER [cf 9] et M.C. PLACE [cf18] ont pour but en particulier :

* de construire un nuage "compromis" entre les nuages associés aux différents tableaux (ce sont dans notre cas les nuages d'individus N_I^j définis par chaque groupe de variables).
* de représenter simultanément ces nuages N_I^j sur des espaces de petite dimension en prenant pour base de départ le nuage compromis.

Nous allons comparer ces méthodes avec la représentation simultanée des nuages N_I^j que nous proposons dont le but est analogue. Ces méthodes vont différer d'une part dans le choix du nuage compromis, d'autre part dans le choix des représentations des N_I^j associés à celle du compromis.

4.3.1 - Choix du nuage compromis

a. Dans la méthode que nous proposons

Dans la méthode que nous proposons, le nuage compromis, que nous appelons nuage moyen, (car il réalise une propriété de moyenne sur les carrés des distances) est tout simplement, à une homothétie près, le nuage associé à l'ensemble de toutes les variables. Chaque groupe de variables j est surpondéré par un coefficient α_j, qui est rappelons-le, l'inverse de la racine carrée de la première valeur propre de l'A.C.P. du groupe de variables j.

La distance dans ce nuage N_I vérifie (cf § 2-3) :

$$d^2(i,i') = \sum_{j \in J} \alpha_j d_j^2(i,i')$$

Le coefficient α_j équilibre le rôle des différents groupes en rendant égale à 1 l'inertie de la projection du nuage $\sqrt{\alpha_j}\, N_I^j$ sur son premier axe d'inertie. Le carré de la distance dans N_I est donc la somme des carrés des distances dans les nuages $\sqrt{\alpha_j}\, N_I^j$ ainsi "normalisés".

b•Dans la méthode STATIS

Dans la méthode STATIS, le nuage compromis est le nuage associé à la première composante principale de l'analyse des opérateurs $W_j D$ évoquée au paragraphe précédent. Les produits scalaires, dans $(\mathbb{R}^I)^2$ entre deux opérateurs $W_j D$ et $W_{j'} D$ étant toujours positifs la première composante principale de ces opérateurs, que nous noterons $\overline{WD}$ est une combinaison linéaire à coefficients positifs des $W_j D$:

$$\overline{WD} = \sum_j \beta_j W_j D \qquad \text{avec} \quad \beta_j \geqslant 0$$

Le nuage associé à $\overline{WD}$ est exactement le nuage associé à l'ensemble de toutes les variables, le groupe j étant surpondéré par β_j.(cf §2)

c•Comparaison

Ces deux solutions sont donc tout à fait comparables.Notre dans le choix des α_j est d'équilibrer le rôle des groupes de variables, celle de H. L'HERMIER DES PLANTES dans le choix des β_j est d'ajuster l'opérateur associé au nuage compromis à l'ensemble des opérateurs associés aux différents groupes de variables.

Le calcul des β_j nécessite l'A.C.P. des opérateurs, mais cette A.C.P. est utilisée aussi pour comparer les groupes de variables. Le calcul des α_j nécessite l'A.C.P. de chaque groupe de variables, mais le résultat de ces A.C.P. est utilisé aussi dans d'autres aspects de notre méthode.

4.3.2 - Représentations simultanées liées au compromis

Les différences entre les méthodes sont beaucoup plus importantes au

niveau de ces représentations que dans le choix des compromis.

a. Dans la méthode que nous proposons

La représentation simultanée des J nuages se déduit de l'analyse en composantes principales du nuage compromis N_I. Aux S premières composantes de cette analyse est associée une représentation de ces nuages dans un espace de dimension S. Rappelons que la composante d'ordre δ de N_I est notée F_δ^I.

La projection de la représentation de N_I^j sur l'axe δ $F_{\delta,j}^I$ se déduit de F_δ^I en appliquant l'opérateur $W_j D$:

$$F_{\delta,j}^I = (1/\lambda_\delta)\, W_j \; D \; F_\delta^I$$

Rappelons les propriétés vérifiées par cette représentation simultanée. Tout d'abord, si l'on regarde ces projections axe par axe :

* $F_{\delta,j}^I$ est la projection orthogonale du nuage exact N_I^j sur un axe.

* Cette projection réalise un compromis entre deux exigences : elle est très correlée avec F_δ^I ; c'est une projection sur un axe de grande inertie du nuage.

Si on regarde globalement ces projections et non plus seulement axe par axe , les représentations simultanées des N_I^j sont des projections orthogonales des N_I^j sur des espaces de dimension S qui réalisent un compromis entre les deux propriétés que nous avions notées :

P_1 Chaque nuage est bien représenté (l'inertie de sa projection est grande).

P_2 Les représentations de ces nuages se ressemblent.

et possède en outre les propriétés :

P_3 Cette représentation contient une représentation du nuage compromis, N_I^* (qui est l'homothétique de N_I dans le rapport $1/J$). Chaque point i^* du nuage N_I^* est situé au centre de gravité des points i^j le représentant dans les nuages N_I^j . Cette propriété facilite beaucoup la comparaison de ces nuages.

P_4 Les projections des variables, des composantes principales de chaque groupe de variables et des opérateurs permettent de préciser et d'analyser les ressemblances et les différences entre les nuages.

Enfin, cette représentation simultanée s'interprète dans l'espace

R^K où l'on peut représenter simultanément les N_I^j mais aussi dans R^I, et dans $(R^I)^*$.

b•Dans la méthode STATIS

La représentation simultanée se déduit aussi de l'analyse en composantes principales du compromis. Nous notons encore F_δ^I la s-ième composante principale du compromis bien qu'il ne s'agisse pas du même compromis que celui du paragraphe précédent.

Nous notons aussi :

$\overline{F}_{\delta,j}^I$: la projection de la représentation de N_I^j sur l'axe δ dans la méthode STATIS.

$\hat{F}_{t,j}^I$: la composante principale d'ordre t de l'analyse du groupe de variables X_j.

$\lambda_{t,j}$: la valeur propre associée à $\hat{F}_{t,j}^I$.

La projection $\overline{F}_{\delta,j}^I$ se déduit de la composante principale F_δ^I du compromis par la formule suivante [cf 16]

$$\overline{F}_{\delta,j}^I = \sum_t \cos\delta\,(F^I, \hat{F}_t^I)\ \hat{F}_{t,j}^I$$

Les composantes principales de X_j sont les vecteurs propres de l'opérateur $W_j D$ et forment une base orthogonale du sous espace E_j engendré par les variables du groupe j. Le carré de leur norme est égal à $\lambda_{t,j}$.

L'application qui permet de déduire la représentation $F_{\delta,j}^I$ du nuage N_I^j de la composante principale du compromis F_δ^I n'est pas la projection orthogonale sur E_j (la projection orthogonale s'obtiendrait en normant les $\hat{F}_{t,j}^I$, i.e. en les divisant par $\sqrt{\lambda_{t,j}}$).

Ce n'est pas non plus l'application $W_j D$ que nous utilisons dans notre méthode (il faudrait multiplier les $\hat{F}_{t,j}^I$ par $\sqrt{\lambda_{t,j}}$). C'est l'opérateur que nous notons $\sqrt{W_j D}$ car son carré est $W_j D$, il a les mêmes vecteurs propres que $W_j D$ et pour valeurs propres les racines carrées de celles de $W_j D$

$$\overline{F}_{\delta,j}^I = \sqrt{W_j D}\ F_t^I$$

Etudions les propriétés de cette représentation.

* Tout d'abord les $\overline{F}_{\delta,j}^I$ sont des projections du nuage N_I^j sur un axe, à une homothétie près, puisque ce sont des vecteurs du sous espace E_j.

* Mais, lorsque l'on considère plusieurs composantes simultanément, i.e. des représentations des N_I^j sur des plans ou des espaces de dimension supérieure à 2, ces représentations ne sont pas des projections des nuages N_I^j. Il n'y a pas d'interprétation géométrique analogue à celle que nous avions dans $\mathbb{R}^K$ puisqu'il ne s'agit pas de projection.

* Pour chaque axe, $\overline{F}_{\delta,j}^I$ peut être considéré comme un compromis entre la ressemblance avec F_δ^I et une bonne qualité de représentation de N_I^j. En effet, la projection linéaire de N_I^j la plus liée à F_δ^I est la projection de F_δ^I sur E_j ; l'application de $\sqrt{W_j D}$ permet d'obtenir des directions de plus grande inertie. Nous appliquions $W_j D$ qui favorisait plus que $\sqrt{W_j D}$ le deuxième aspect et qui avait une propriété optimale (rendre maximum le produit scalaire de F_δ^I et $F_{\delta,j}^I$) . Mais surtout, le compromis entre les deux propriétés (P_1) et (P_2) (la ressemblance entre les représentations simultanées des nuages et leur qualité de représentation) était réalisé globalement et non pas seulement axe par axe.

* Le nuage compromis n'est généralement pas au centre de gravité des représentations des intrastrutures. Il se trouve au centre de gravité si l'opérateur $W_j D$ est égal à la projection sur E_j . Dans ce cas très particulier, $\sqrt{W_j D}$ et $W_j D$ sont confondus et la propriété vérifiée par notre méthode l'est aussi ici.

* Si on prend toutes les composantes du nuage compromis, on obtient une représentation exacte des nuages N_I^j .Cette propriété n'était pas vérifiée pour nous. (Dans $\mathbb{R}^K$ le sous espace engendré par le nuage moyen N_I ne contient pas tous les nuages N_I^j ; en projetant les N_I^j sur les axes d'inertie de N_I , leur qualité de représentation n'est pas égale à 1).

c. Dans les variantes de la méthode STATIS

Y. ESCOUFIER 9 et M.C. PLACE 18 proposent de projeter les premières composantes principales du groupe X_j sur le sous-espace engendre par les premières composantes principales du compromis.

On sait que le compromis peut être considéré comme un nuage associé à l'ensemble de toutes les variables (surpondérées par les β_j). Le sous espace engendré par ses composantes principales est donc le sous espace engendré par toutes les variables et contient les composantes principales de chaque groupe.

Si l'on retenait toutes les composantes du compromis (ce qui est peu réaliste), cette méthode reviendrait à superposer les composantes principales de tous les groupes.

Sinon, la projection de ces composantes $F^I_{t,j}$ a toutes chances de ne pas appartenir au sous espace E_j et ne pourra pas être considérée comme une projection du nuage N^j_I sur un axe.

d Comparaison

Nous avions dégagé dès l'introduction de ce travail les propriétés qui nous paraissaient souhaitables dans une bonne représentation simultanée des nuages.

La première, (que cette représentation soit une projection orthogonale des nuages) est vérifiée seulement sur chaque axe par la méthode STATIS et pas du tout dans la variante. Elle est vérifiée dans notre méthode.

Les propriétés suivantes (de compromis entre qualité de représentation et ressemblance entre les projections) ne peuvent s'exprimer qu'axe par axe pour STATIS. Elles ne peuvent s'exprimer dans la variante puisque les représentations approchées des nuages ne sont pas des projections.

Le nuage moyen est situé au centre de gravité des différents nuages dans notre méthode seulement.

4.4. Le modèle INDSCAL

La solution que nous proposons pour le modèle INDSCAL a l'avantage d'être très simple du point de vue des calculs et ne pose aucun problème de convergence d'algorithme puisqu'elle est basée sur la diagonalisation d'un opérateur symétrique.

Elle s'interprète géométriquement dans l'espace R^{I^2} où les poids des différents facteurs pour chaque groupe de variables sont tout simplement les coordonnées de l'opérateur associé au groupe (i.e. la matrice des produits scalaires) sur un système orthonormé engendré par les facteurs. Elle s'interprète aussi dans l'espace R^K où les images de chaque nuage apparaissent comme des approximations des projections de ces nuages.

Ceci permet d'introduire des aides à l'interprétation des résultats et des mesures très précises, facteur par facteur, de l'approximation donnée par le modèle.

Si le modèle est exact, la solution juste est obtenue. Des critères sont maximisés. Les facteurs communs obtenus, les poids affectés par un groupe de variables minimisent la norme de la différence entre la matrice des produits scalaires réels et celle du modèle. La matrice du modèle est alors une projection dans $(R^I)^2$ de la matrice réelle. Les facteurs sont orthogonaux deux à deux et maximisent la somme (sur l'ensemble des groupes) des normes de ces projections.

4.5. Les méthodes d'analyse Procustéenne [cf 2,10, 21]

Le but de ces méthodes est de faciliter la comparaison de deux ou plusieurs nuages de n points homologues situés dans le même espace euclidien. Ces nuages, même s'ils se ressemblent, ne sont pas, a priori, orientés de la même façon. Les techniques d'analyse procustéenne consistent à chercher des déplacements de chacun des nuages qui rapprochent les points homologues

Les déplacements autorisés sont des isométries laissant fixe l'origine i.e. des transformations orthogonales. Avant de chercher ces déplacements, les nuages sont centrés (pour faire coïncider leurs centres de gravité) et normalisés en donnant à chaque nuage la même inertie.

Dans le cas de deux nuages seulement, un nuage est déplacé de manière à minimiser la somme des carrés des distances entre les points homologues des deux nuages (cette somme est l'écart de Procuste entre les deux nuages). La solution est assez simple. La généralisation à plus de deux nuages [cf.21] nécessite des calculs plus complexes. Les nuages étant ainsi orientés de manière analogue, on pourra les représenter dans des espaces de petite dimension par une analyse en composantes principales des points de tous les nuages ou des n points d'un nuage moyen.

Une technique de ce type pourrait être utilisée pour la représentation simultanée des J nuages d'individus associés aux J groupes de variables basée sur la représentation (artificielle) de ces nuages dans le même espace R^K proposée au § 2.3.2.

Comparons cette démarche à celle que nous proposons.

. La normalisation des nuages du § 2.3.3. ne rend pas égales les inerties totales des différents nuages, car nous supposons que des directions de dispersion peuvent exister dans certains nuages et non dans d'autres. La normalisation par l'inertie totale suppose implicitement que les structures générales des nuages sont globalement analogues.

. Une démarche basée sur l'analyse procustéenne opèrerait en deux temps : orientation des différents nuages, puis réduction des dimensions de ces nuages. Ainsi que nous l'avons déjà indiqué au § 1.4., pour dégager les structures communes des nuages nous avons préféré chercher simultanément à orienter et réduire les nuages. En effet, les directions de grande inertie du nuage moyen ou de l'ensemble des nuages ne sont pas forcément des directions où les projections des nuages se ressemblent même si ces nuages ont été réorientés correctement les uns par rapport aux autres.

4.6. Le cas de groupes réduits à une seule variable numérique

Pour ce cas limite, nous nous contenterons de constater que la méthode proposée se réduit à une analyse en composante principale normée, ce qui nous paraît tout à fait satisfaisant. Le seul éclairage nouveau apporté par la méthode en question est l'interprétation géométrique dans $(R^I)^2$ des contributions à l'inertie de chaque variable à une composante. Mais ceci ne paraît pas présenter ici un grand intérêt.

4-7. Cas des variables qualitatives

Rappelons qu'une variable qualitative est une partition sur l'ensemble des individus et qu'elle est représentée par les variables indicatrices des classes de cette partition. Rappelons aussi (cf. § 2-.5.1.d) que ces variables indicatrices sont affectées d'un poids qui rend l'opérateur associé $W_j D$ égal à l'opérateur de projection sur le sous espace engendré par ces variables.

Nous allons étudier deux situations. Dans la première, chaque groupe de variables est constitué par une seule variable qualitative. Nous obtiendrons alors l'analyse de correspondances multiples. Dans la seconde, un groupe de variables est constitué par un nombre quelconque de variables qualitatives, ce qui permet de comparer des groupes de variables qualitatives. Nous évoquerons aussi le cas où apparaissent des groupes de variables qualitatives et de variables numériques.

4.7.1. - Groupesd'une variable. Analyse des correspondances multiples.

Rappelons que :

l'analyse des correspondances multiples est l'analyse des correspondances appliquée au tableau disjonctif complet dont les colonnes sont les variables indicatrices δ_{ij} des classes de toutes les partitions.

Nous montrons que l'analyse des correspondances peut être considérée comme un cas particulier de la méthode proposée. Nous interprétons ses résultats dans toutes les optiques que nous avons étudiées.

4.7.1.a - Analyse multicanonique

Puisque nous avons imposé aux variables indicatrices des poids rendant l'opérateur associé à chaque variable égal à l'opérateur de projection, notre méthode se confond dans ce cas particulier avec l'analyse multicanonique au sens de CARROLL. Ceci était d'ailleurs l'un des buts poursuivis (cf § 1-6)

Rappelons brièvement que l'analyse des correspondances multiples peut aussi être considérée comme un analyse multicanonique [cf.8 et 19]. Tout d'abord, la marge sur I d'un tableau disjonctif complet étant constante, la métrique de R^I est, à un coefficient près, la métrique identité. Dans l'analyse multicanonique des groupes de variables indicatrices, c'est aussi le cas pour R^I, si les individus ont le même poids.

Dans l'analyse des correspondances, la modalité k est représentée par son profil δ_{ik}/δ_k qui est proportionnel à la variable indicatrice δ_{ik} : en outre, on donne le poids δ_k ce qui rend son inertie égale à 1/J.

Les variables indicatrices d'une même partition étant orthogonales deux à deux l'inertie de la projection des modalités de la variable qualitative j sur un axe u de R^I est égale à $(\cos^2\theta j)/J$, où θ_j est l'angle entre u et le sous espace engendré par les modalités de j.

Le premier facteur sur I de l'analyse des correspondances multiples rend maximum l'inertie des projections des modalités, donc la somme des $\cos^2\theta_j$. Il est donc confondu avec la première variable générale de l'analyse multicanonique. Il en est de même pour les autres facteurs qui sont orthogonaux au premier.

Les valeurs propres de l'analyse des correspondances multiples sont égales à l'inertie du nuage projeté donc à $\sum_j (\cos^2\theta j)/J$.

Les variables canoniques associées à un facteur F_α^I sont les projections de F_α^I sur les sous espaces E_j. Ces projections se calculent ici très simplement car les variables indicatrices δ_k $(k\varepsilon K_j)$ de la partition définie par j forment une base orthogonale de E . On a donc :

$$P_j(F_\alpha^I) = \sum_{k\varepsilon K_j} \langle F_\alpha^I \delta_k \rangle \delta_k / \|\delta_k\|^2$$

Cette fonction ne dépend que de la modalité k de l'individu, on peut donc la considérer comme une fonction sur K_j. Elle s'écrit

$$F_{\alpha,j}^I(k) = \sum_{i\varepsilon I} \delta_{ik} F_\alpha^I(i) / \delta_k$$

On reconnait, à une homothétie près, la restriction à K_j du facteur sur K de l'analyse des correspondances multiples.

Les facteurs sur I de l'analyse des correspondances multiples sont les variables générales de l'analyse multicanonique.
Les valeurs propres valent $(1/J) \sum_j \cos^2\theta_j$.
Les facteurs sur K, *restreints à* K_j *sont les variables canoniques.*

4.7.1.b. - Pondération des groupes

Nous proposions de pondérer les groupes de variables par l'inverse de la racine carrée de la première valeur propre de $W_j D$. Ici, toutes les valeurs propres de $W_j D$ sont égales à 1 ou : il n'y a pas besoin de surpondérer les groupes de variables ce qui permet de rester dans le cadre de l'analyse des correspondances multiples.

4.7.1.c. - Représentation simultanée des nuages N_I^j

Dans le cas des variables qualitatives, les nuages N_I^j comprennent peu de points car les individus ayant choisi la même modalité k de K_j sont tous confondus. La distance entre deux individus i et i' ayant choisi les modalités différentes k et k' vaut

$$d_j^2(i,i') = \frac{1}{\delta_k} + \frac{1}{\delta_{k'}}$$

Dans le nuage moyen, nuage des individus défini par le tableau disjonctif complet, le carré de la distance entre deux individus est la somme sur j des d_j^2.

Les facteurs F_α^I peuvent être considérés comme les projections du nuage moyen, compromis entre ces J nuages.

On obtient alors, en appliquant les formules des paragraphes 2-1 les projections $F_{\alpha,j}^I$ des nuages N_I^j, associés à F_α^I. Il est facile de vérifier que ces projections sont à $1/\sqrt{\lambda}$ près les restrictions des facteurs sur K aux K_j.

Donc si dans l'analyse des correspondances multiples, on représente le facteur sur KG_α^K *de norme 1 (est non de norme* $\sqrt{\lambda}$ *comme on le fait classiquement) nous obtiendrons la représentation simultanée des nuages* N_I^j *que nous proposons :* $F_\alpha(i)$ *est le point représentant* i *dans le nuage moyen ; pour* $k \in K_j$ $G_\alpha(k)$ *représente, vus dans le nuage* N_I^j *les individus* i *qui ont la modalité* k *de* K_j ; $F_\alpha(i)$ *est alors au centre de gravité des modalités qu'il prend pour les diverses variables, i.e. de ses représentations dans les différents nuages* N_I^j.

4.7.1.d. - Comparaisons globale des variables

Nous avons déjà proposé dans [7] une représentation des variables qualitatives dans l'analyse des correspondances multiples. La technique proposée ici apparait comme une généralisation de celle évoquée ci-dessus à des groupes de variables quelconques.

4.7.1.e. - Projection des composantes principales des groupes.

Cela ne présente pas d'intérêt ici puisque les composantes principales des groupes ne sont pas déterminées, la valeur propre 1 étant multiple.

4.7.1.f. - Les indices d'aide à l'interprétation.

Les indices de ressemblance entre les projections des N_I^j peuvent être utilisés,

L'indice global de ressemblance entre les projections : inertie inter/inertie totale (i.e. inertie de la projection du nuage moyen/Σinertie des projections des N_I^j) se réduit à l'inertie inter (i.e. la valeur propre) car l'inertie totale vaut toujours 1. En effet, plaçons-nous dans R^K, les différents nuages N_I^j sont situés dans les sous-espaces R^{Kj} orthogonaux deux à deux et se projettent sur un vecteur u de R^K.

L'inertie de N_I^j vaut 1 dans toutes les directions du sous-espace qu'il engendre. L'inertie et sa projection sur u vaut donc $\cos^2 H_j$ où H_j est l'angle entre u et R^{Kj}, or, la somme des $\cos^2 H_j$ vaut 1 puisque les R^{Kj} sont orthogonaux. Nous retrouvons directement ce résultat sur les facteurs obtenus puisque la représentation simultanée des J nuages est donnée par le facteur $G^J{}_j$ de norme 1.

La contribution d'une variable à l'inertie λ d'un facteur est la somme des contributions de ses modalités. Elle vaut $(\cos^2\theta j)/\lambda$. Elle est égale aussi, comme toujours à $\cos^2 H_j$ (angle dans R^K). On a donc $\cos\theta_j = \lambda\cos^2 H_j$.

L'inertie totale d'un nuage N_I^j (ou du nuage des modalités de j) est égale au nombre de ces modalités diminué d'une unité puisque l'on prend un nuage centré. La qualité de représentation de la variable qualitative j par le facteur d'ordre α est donc égale dans R^I à $\cos^2\theta_j/(\text{card } K_j - 1)$. Dans R^K, c'est la même valeur multipliée par la valeur propre λ. En effet l'inertie de la projection de N_I^j sur l'axe u est égale à $\cos^2 H_j$. Dans $(R^I)^2$ le carré de la norme de W_j est

encore égal à card(K_j-1). L'inertie de sa projection est $\cos^4 H_j$, sa qualité de représentation est donc $\lambda^2 \cos^4 \theta_j / (\text{card } K_j - 1)$

4.7.2.- Groupe de plusieurs variables qualitatives

La méthode proposée peut s'appliquer à l'étude de plusieurs groupes de variables qualitatives. L'analyse de chaque groupe est exactement l'analyse des correspondances des tableaux disjonctifs complets. Chaque groupe est ensuite surpondéré par l'inverse de la racine carréede la première valeur propre de son analyse. A cause de cette pondération l'analyse globale n'est plus tout à fait une analyse des correspondances multiples. Les résultats sont ceux du cas général de groupes de variables numériques ; les particularités du paragraphe précédent disparaissent dès que plusieurs variables qualitatives interviennent puisque les valeurs propres non nulles des opérateurs associés ne sont plus égales entre elles.

Un ensemble de variables qualitatives apparait donc tout à fait analogue à un groupe de variables numériques. Mélanger ces deux types de groupes de variables ne pose pas de problèmes de fond. Les facteurs du nuage moyen sont les composantes principales de composantes de chaque groupe (cf. §3.2.4.). Ce sont les composantes principales des facteurs des analyses de correspondances des tableaux disjonctifs complets des groupes de variables qualitatives ; et des composantes principales des groupes de variables numériques. On retrouve une technique utilisée quelquefois pour traiter simultanément variables numériques et variables qualitatives. La pondération de chaque groupe que nous proposons reste intéressante ici aussi.

5 BIBLIOGRAPHIE

[1] BENZECRI J.P. et coll. (1973)
L'analyse des données.
DUNOD, Paris.

[2] BOURGEOIS Ph. (1980)
Recherche du déplacement minimisant la distance entre deux configurations de points indicées par le même ensemble.
Thèse de 3ème cycle, PARIS VI.

[3] CAILLEZ F., PAGES J.P. (1976)
Introduction à l'analyse des données.
SMASH, Paris.

[4] CARROL J.D. et CHANG J.J. (1970)
Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart Young" decomposition.
Psychometrika, vol. 35 n° 3, p. 283-319.

[5] CARROL J.D. (1968)
A generalization of canonical correlation analysis to three or more sets of variables.
Proceedings of the 76th annual convention of the American Psychological association, p. 227-228.

[6] CAZES P., ESCOUFIER Y., PAGES J.P. (1976)
Opérateurs et analyse des tableaux à plus de deux dimensions.
Cahiers du Buro n° 25.

[7] ESCOFIER B. (1979)
Une représentation des variables dans l'analyse des correspondances multiples.
Revue de Statistiques Appliquées, 1979, n° 4.

[8] ESCOFIER B. (1980)
Le traitement des variables qualitatives.
Cours multigraphié, Université de Rennes 1.

[9] ESCOUFIER Y. (1980)
L'analyse conjointe de plusieurs matrices
Biométrie et temps
Société Française de Biométrie

[10] GOWER J.C. (1975)
Generalized procustres analysis.
Psychometrika, vol. 40 n° 1, p. 33-51.

[11] HORST P.(1961)
Relations among m sets of measures.
Psychometrika, vol. 26 n° 2, p. 129-149.

[12] HOTELLING H. (1936)
Relations between two sets of variables.
Biometrika, n° 28, p. 277-321.

[13] KETTENRING J.R. (1976)
Canonical analysis of several sets of variables.
Biometrika, vol. 58 n° 3, p. 433-451.

[14] KOBILINSKY A. (1977)
Propriétés et utilisation de l'analyse multicanonique par la méthode de Carroll.
Analyse des données et Informatique, IRIA Rocquencourt.

[15] LEBART L., MORINEAU A., TABARD N. (1977)
Techniques de la description statistique : méthodes et logiciels pour l'analyse des grands tableaux.
DUNOD, Paris.

[16] L'HERMIER DES PLANTES H. (1976)
Structuration des tableaux à trois indices de la statistique.
Thèse de 3ème cycle, Université de Montpellier.

[17] PAGES J.P., CAILLEZ F., ESCOUFIER Y.
Analyse factorielle : un peu d'histoire et de géométrie.

[18] PLACE M.C. (1980)
Contribution algorithmique à la mise en oeuvre de la méthode STATIS.
Thèse de 3ème cycle, Université des Sciences et Techniques du Languedoc, Montpellier.

[19] SAPORTA G. (1975)
Liaisons entre plusieurs ensembles de variables et codage de données qualitatives.
Thèse de 3ème cycle, Paris VI.

[20] TAKANE Y. (1977)
Nonmetric individual differences multidimensional scaling : an alternating least squares method with optimal scaling features.
Psychometrika, vol. 42, n° 1.

[21] TEN BERGE (1977)
Orthogonal procustres rotation for two or more matrices.
Psychometrika, vol. 42 n° 2, p. 267-276.

[22] WOLLENBERG DEN M.L. (1977)
Redundancy analysis : an alternative for canonical correlation analysis.
Psychometrika, vol. 42 n° 2.

Liste des Publications Internes IRISA

PI 139 **Efficacité des algorithmes récursifs en présence de systèmes non-stationnaires.**
A. Benveniste, G. Ruget , 35 pages ; *Août 1980*

PI 140 **Structures de communication extensibles**
P. Le Guernic, M. Raynal , 60 pages ; *Octobre 1980*

PI 141 **Comparaison de tableaux de fréquence**
B. Escoffier , 16 pages ; *Octobre 1980*

PI 142 **Un lemme général de stabilité pour la commande adaptative en déterministe de systèmes non nécessairement à minimum de phase]:**
Cl. Samson , 40 pages ; *Novembre 1980*

PI 143 **Détection, Estimation de l'orientation et saisie d'une cible mobile par proximétrie optique**
B. Espiau , 142 pages ; *Janvier 1981*

PI 144 **Une contribution à l'étude de l'impact de l'informatique sur les organisations**
L. Breton, A. Prod'homme, J. Villard , 58 pages ; *Décembre 1980*

PI 145 **Rupture de modèles statistiques**
M. Basseville, A. Benveniste , 130 pages ; *Mars 1981*

PI 146 **Traitement des questionnaires avec non réponse, analyse des correspondances avec marge modifiée et analyse multicanonique avec contrainte**
B. Escoffier , 38 pages ; *Mars 1981*

PI 147 **Deux files d'attente à capacité limitée en tandem**
J. Pellaumail, J. Boyer , 19 pages ; *Juillet 1981*

PI 148 **Programme de classification hiérarchique : 1) Méthode de la vraisemblance des liens, 2) Méthode de la variance expliquée**
I.C. Lerman , 113 pages ; *Juin 1981*

PI 149 **Convergence des méthodes de commande adaptative en présence de perturbations aléatoires**
J.J. Fuchs , 46 pages ; *Juillet 1981*

PI 150 **Construction automatique et évaluation d'un graphe d'«implication» issu de données binaires, dans le cadre de la didactique des mathématiques**
H. Rostam , 112 pages ; *Juin 1981*

PI 151 **Réalisation d'un outil d'évaluation de mécanismes de détection de pannes]-]Projet Pilote SURF**
B. Decouty, G. Michel, C. Wagner, Y. Crouzet , 59 pages ; *Juillet 1981*

PI 152 **Règle maximale**
J. Pellaumail , 18 pages ; *Septembre 1981*

PI 153 **Corrélation partielle dans le cas « qualitatif »**
I.C. Lerman , 125 pages ; *Octobre 1981*

PI 154 **Stability analysis of adaptively controlled not-necessarily minimum phase systems with disturbances**
Cl. Samson , 40 pages ; *Octobre 1981*

PI 155 **Analyses d'opinions d'instituteurs à l'égard de l'appropriation des nombres naturels par les élèves de cycle préparatoire**
R. Gras , 37 pages ; *Octobre 1981*

PI 156 **Récursion induction principle revisited**
G. Boudol, L. Kott , 49 pages ; *Décembre 1981*

PI 157 **Loi d'une variable aléatoire à valeur $\mathbf{R}^+$** réalisant le minimum des moments d'ordre supérieur à deux lorsque les deux premiers sont fixés
M.Kowalowka, R. Marie , 8 pages ; *Décembre 1981*

PI 158 **Réalisations stochastiques de signaux non stationnaires, et identification sur un seul échantillon**
A. Benveniste J.J. Fuchs , 33 pages ; *Mars 1982*

PI 159 **Méthode d'interprétation d'une classification hiérarchique d'attributs-modalites pour l'«explication» d'une variable ; application à la recherche de seuil critique de la tension artérielle systolique et des indicateurs de risque cardiovasculaire**
B. Tallur , 34 pages ; *Janvier 1982*

PI 160 **Probabilite stationnaire d'un réseau de files d'attente multiclasse à serveur central et à routages dépendant de l'état**
L.M. Le Ny , 18 pages ; *Janvier 1982*

PI 161 **Détection séqentielle de changements brusques des caractéristiques spectrales d'un signal numérique**
M. Basseville, A. Benveniste , pages ; *Mars 1982*

PI 162 **Actes regroupés des journées de Classification de Toulouse (Mai 1980), et de Nancy (Juin 1981)**
I.C. Lerman , 304 pages ;

PI 163 **Modélisation et Identification des caractéristiques d'une structure vibratoire : un problème de réalisation stochastique d'un grand système non stationnaire**
M. Prévosto, A. Benveniste, B. Barnouin , 46 pages ; *Mars 182*

PI 164 **An enlarged definition and complete axiomatization of observational congruence of finite processes**
Ph. Darondeau , 45 pages ; *Avril 1982*

PI 165 **Accès vidéotex à une banque de données médicales**
A. Chauffaut, M. Dragone, R. Rivoire, J.M. Roger , 25 pages ; *Mai 1982*

PI 166 **Comparaison de groupes de variables définies sur le même ensemble d'individus**
B. Escofier, J. Pages , 115 pages ; *Mai 1982*

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique